RCAAP
Repositório Científico de Acesso Aberto de Portugal

OS REPOSITÓRIOS DE DADOS CIENTÍFICOS: ESTADO DA ARTE

JULHO 2010

MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
UMIC Agência para a Sociedade do Conhecimento
FCCN Fundação para a Computação Científica Nacional
Universidade do Minho
Financiado por POS_Conhecimento Programa Operacional Sociedade do Conhecimento
União Europeia FEDER

Ficha Técnica:

PROJECTO RCAAP
D24 - RELATÓRIO

Os Repositórios de Dados Científicos: Estado da arte

Autoria: Grupo de trabalho conjunto da Universidade do Minho e da Universidade do Porto:

Universidade do Minho: Eloy Rodrigues e Ricardo Saraiva (com colaborações de Cristina Marques Gomes e José Carvalho)

Universidade do Porto: Cristina Ribeiro e Eugénia Matos Fernandes

Versão: 1.0

Data de criação: 17 de Maio de 2010

Última actualização: 4 de Julho de 2010

MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR · UMIC Agência para a Sociedade do Conhecimento · FCCN Fundação para a Computação Científica Nacional · Universidade do Minho

Financiado por POS_Conhecimento Programa Operacional Sociedade do Conhecimento · União Europeia FEDER

SUMÁRIO

SUMÁRIO EXECUTIVO

O presente estudo está enquadrado nas actividades de 2010 do projecto Repositório Científico de Acesso Aberto de Portugal (RCAAP) e assinala o início da sua intervenção no domínio da curadoria dos dados resultantes das actividades de investigação, da sua organização em repositórios de dados científicos e do seu acesso. Para além de recolher informação actualizada sobre o tema e as iniciativas mais relevantes relacionadas com a gestão e acesso aos dados científicos através de repositórios, o objectivo deste documento é também o de informar e orientar o desenvolvimento de um projecto-piloto de repositório de dados científicos que está também previsto no plano de trabalho do projecto RCAAP para 2010.

O documento inicia-se com uma introdução, que contextualiza a crescente visibilidade dos temas relacionados com a curadoria e a partilha dos dados científicos e explicita as escolhas terminológicas que foram realizadas para os conceitos mais comuns na literatura neste domínio.

Na segunda secção do estudo intitulada: *"Dados científicos e repositórios de dados"*, é apresentado um quadro actual dos repositórios de dados científicos, referindo o seu enquadramento, a sua origem e a sua evolução. A necessidade de conjugar a dimensão institucional (muito ampla e multidisciplinar no caso das universidades) com a dimensão disciplinar (com os seus requisitos específicos) é identificada como um dos principais desafios à utilização dos repositórios institucionais como componente da infra-estrutura global de curadoria dos dados científicos. Nesta secção são ainda apresentadas e descritas as principais tecnologias, plataformas e normas de metadados utilizadas neste domínio.

Na terceira secção designada: *"Repositórios, partilha e curadoria de dados"*, o relatório prossegue com uma identificação dos principais actores, problemas, desafios, soluções e benefícios relacionados com o acesso e a gestão de dados científicos através de repositórios. Constata-se que a tomada de consciência da necessidade do armazenamento e da preservação de dados científicos em repositórios criados e mantidos para esse efeito constitui um processo ainda em curso, com diferentes estádios de maturidade a nível internacional e que se afigura como indispensável a aproximação entre os investigadores e as instituições que gerem repositórios para alojamento, preservação e acesso a dados científicos. Nesta secção são ainda revistos os aspectos políticos, legais e éticos associados ao acesso e reutilização dos dados científicos para além do contexto inicial em que foram recolhidos.

Nas conclusões, que constituem a última secção do documento, constata-se que, apesar do crescente interesse que o tema vem despertando, com a multiplicação de actividades, iniciativas e projectos nos últimos anos, a curadoria e partilha de dados científicos é uma área "jovem", ainda em formação e consolidação. Esta circunstância constitui uma oportunidade para a investigação e desenvolvimento de novos serviços e tecnologias, mas simultaneamente um desafio e um risco para o funcionamento de serviços de qualidade profissional.

Reconhecendo que a curadoria, para ser verdadeiramente efectiva e sustentável, exige a participação de todas as partes envolvidas na produção dos dados científicos, o texto termina com a apresentação de acções e orientações a serem desenvolvidas pelos investigadores, as instituições de investigação, os organismos de financiamento e os responsáveis de repositórios, de que aqui se destacam as principais.

Investigadores:

- Incluir a curadoria, e eventual partilha, dos dados científicos, no processo de planeamento da investigação, analisando e identificando as soluções e as práticas adequadas, desejavelmente produzindo um plano de curadoria.
- Colaborar e cooperar com os serviços e projectos de curadoria de dados relevantes para o seu contexto, disciplinar e institucional, a fim de conhecer, utilizar e promover as boas práticas neste domínio.
- Divulgar e partilhar os dados científicos que produzam, tão cedo e tão amplamente quanto possível em cada caso, sem prejuízo dos seus próprios interesses ou dos constrangimentos legais e éticos que possam existir.

Instituições de investigação:

- Realizar recenseamento e diagnóstico da situação existente no que diz respeito aos dados científicos em cada instituição.
- Disponibilizar, de forma autónoma ou colaborativa com outras instituições, infra-estruturas e serviços para a curadoria dos dados científicos, seja através dos repositórios institucionais já existentes, seja através de repositórios de dados especificamente criados para esse efeito.
- Atribuir a competência pela área da curadoria de dados científicos a uma unidade organizacional da instituição, devidamente apetrechada com os necessários recursos financeiros e humanos.

- Incentivar os investigadores a preocupar-se com a curadoria dos dados que produzem, e definir políticas institucionais que induzam o depósito dos dados científicos em repositórios institucionalmente adequados e estimulem a partilha dos dados depositados tão cedo e de forma tão ampla quanto seja possível.
- Avaliar e identificar as necessidades de formação de técnicos de curadoria de dados para as diferentes áreas científicas e disciplinares.

Organismos financiadores de investigação:

- Definir políticas que exijam, ou pelo menos valorizem significativamente na avaliação para financiamento, a existência de um plano de curadoria de dados em todos os projectos de investigação.
- Definir políticas e procedimentos que exijam o depósito dos dados científicos em repositórios sustentáveis e o acesso aberto a esses dados sempre e logo que possível.
- Considerar como elegíveis para financiamento, nos projectos que financiam, as eventuais despesas dos investigadores com actividades de curadoria e partilha de dados.
- Disponibilizar financiamento específico para a realização de projectos de investigação e desenvolvimento no domínio da curadoria de dados científicos.

Responsáveis por repositórios:

- Assegurar que os repositórios colocados à disposição dos investigadores sejam infra-estruturas robustas e fiáveis do ponto de vista tecnológico e da segurança da informação.
- Disponibilizar aos investigadores serviços e ferramentas de apoio à curadoria e partilha de dados.
- Recolher e disponibilizar informação sobre acesso e utilização dos conjuntos de dados que gerem.
- Acompanhar as iniciativas internacionais relacionadas com a criação, gestão e manutenção de repositórios de dados, recolhendo a experiência de projectos pioneiros e aprendendo com os seus resultados.

Finalmente, considerando a existência de pontos comuns, para além de importantes diferenças, entre o arquivo e a disponibilização de publicações e a curadoria e disponibilização de dados, sugere-se que a experiência acumulada pelo projecto RCAAP no domínio dos repositórios institucionais pode constituir uma base para o lançamento de iniciativas na área dos repositórios de dados científicos.

1 – INTRODUÇÃO

O presente documento foi realizado no âmbito do projecto Repositório Científico de Acesso Aberto de Portugal (RCAAP)[1]. A iniciativa RCAAP visa aumentar a visibilidade, acessibilidade e difusão dos resultados da actividade académica e de investigação científica nacional, facilitar o acesso à informação sobre a produção científica nacional em regime de acesso aberto, bem como integrar Portugal num conjunto de iniciativas internacionais neste domínio.

Até ao final de 2009 as actividades do projecto RCAAP focaram-se exclusivamente nos repositórios de literatura científica. O plano de trabalho para 2010 assinala o início da intervenção do projecto RCAAP no domínio do acesso e curadoria dos dados resultantes das actividades de investigação e dos repositórios de dados científicos.

Os dois resultados esperados da actividade do RCAAP nesta área, em 2010, são o presente relatório de estado da arte e um projecto-piloto. Para além de recolher informação actualizada sobre o tema e as iniciativas mais relevantes relacionadas com a gestão e acesso aos dados científicos através de repositórios, o objectivo deste documento é também o de informar e orientar o desenvolvimento desse projecto-piloto de repositório de dados científicos.

Existem duas grandes áreas de requisitos no domínio dos dados científicos. A primeira está relacionada com as infra-estruturas (tais como serviços, sistemas, normas e protocolos) necessárias para garantir a recolha, preservação e acesso aos dados. A segunda tem a ver com os aspectos políticos, legais e éticos, associados ao acesso e reutilização dos dados científicos para além do contexto inicial em que foram recolhidos.

O alargamento da actividade do projecto RCAAP a este domínio corresponde ao reconhecimento da crescente importância das questões relativas aos dados resultantes das actividades de investigação no sistema de comunicação científica, e mesmo no funcionamento global da ciência. De facto, na última década, os temas relacionados com a gestão, o acesso e reutilização dos dados científicos, e em especial dos resultantes de actividades de investigação financiadas com fundos públicos, adquiriram visibilidade e

[1] O projecto RCAAP é uma iniciativa da UMIC – Agência para a Sociedade do Conhecimento, IP concretizada pela FCCN – Fundação para a Computação Científica Nacional, disponibilizando mais um serviço avançado sobre a Rede Ciência, Tecnologia e Sociedade (RCTS) gerida pela FCCN. A execução do projecto conta ainda com a participação científica e técnica da Universidade do Minho. Toda a informação do projecto pode ser consultada em: http://projecto.rcaap.pt/.

atraíram acrescida atenção, quer no seio de comunidade científica, quer na esfera política ao nível internacional.

Do lado das comunidades científicas, está a crescer a sensibilidade para as consequências da verdadeira explosão na produção de dados científicos, quer pelo crescimento global das actividades de investigação, quer como consequência dos novos métodos e instrumentos de pesquisa e registo que originam cada vez maiores volumes de dados. Para além da produção, recolha, análise e interpretação dos dados, onde naturalmente os investigadores concentram o seu trabalho, tem vindo a afirmar-se uma outra área de intervenção, relacionada com a curadoria dos dados científicos.

O entendimento da necessidade de gerir o acesso e a utilização dos dados produzidos ou recolhidos no quadro das actividades de investigação, garantindo a sua preservação, parece ter sido mais precoce e mais profundo em áreas como a astronomia e a climatologia, onde o trabalho de investigação é baseado na análise de dados recolhidos de forma distribuída, mas tem vindo a alargar-se a outros domínios.

Do lado dos decisores políticos tem havido também um número crescente de declarações e iniciativas, quer a nível nacional, quer no contexto europeu e em organizações internacionais. Um dos primeiros e mais relevantes sinais desta realidade foi a Declaração sobre o acesso a dados científicos com financiamento público, aprovada pelos representantes ministeriais de 34 países (incluindo Portugal) da OCDE em 30 de Janeiro de 2004[2].

A Declaração da OCDE, reconhecendo que o acesso livre e o uso irrestrito dos dados promovem o progresso científico e maximizam o retorno do investimento público nas actividades de recolha de dados, e que restrições indevidas ao acesso e utilização dos dados científicos podem diminuir a qualidade e eficiência da investigação científica e inovação, afirma a vontade de trabalhar para o estabelecimento de regimes de acesso aos dados científicos resultantes de financiamento público. No seguimento desta declaração de 2004, a OCDE aprovou em 2006 um documento de Princípios e Directrizes para o Acesso aos dados científicos resultantes de financiamento público, publicado já em 2007[3] e que constitui um dos documentos de referência neste domínio.

---

[2] OECD. *Declaration on Access to Research Data From Public Funding*, Paris, 2004. Disponível em: http://www.oecd.org/document/15/0,3343,en_2649_34487_25998799_1_1_1_1,00.html [Consultado em 26 de Abril 2010].

[3] *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris, 2007. Disponível em: http://www.oecd.org/dataoecd/9/61/38500813.pdf [consultado em: 26 de Abril de 2010].

A questão do acesso aberto aos dados científicos tem conquistado também apoio crescente nas comunidades científicas, multiplicando-se as iniciativas, projectos e declarações. Não deixando de estar relacionado com o ambiente geral e outros movimentos de "abertura", na sociedade e na ciência em particular, como os do *Open Source* e *Open Access*, o "movimento" *Open Data* tem origem, dinâmica e "agenda" próprias.

Na redacção deste relatório foi necessário tomar decisões quanto à forma e às definições de vários conceitos comuns na literatura sobre este tema, que geralmente se exprimem em língua inglesa, no sentido de garantir a uniformidade e inteligibilidade da sua utilização. Assim, por exemplo, decidiu-se adoptar a expressão **dados científicos** para ***research data***, **conjunto de dados** para ***dataset*** e **curadoria de dados** para ***data curation***. Uma breve explicitação destes conceitos é apresentada na Secção 2.1, e incluiu-se no final um glossário dos termos usados ao longo do relatório.

Para além desta Introdução, o relatório está estruturado em mais 3 secções e um anexo. A Secção 2 apresenta uma panorâmica dos repositórios de dados científicos, contextualizando o seu enquadramento, a sua origem, evolução e situação actual. A Secção 3 identifica os principais actores, problemas, desafios e soluções relacionados com o acesso e a gestão de dados científicos através de repositórios. Finalmente, a Secção 4 apresenta algumas conclusões e sugestões de acção para as diferentes partes interessadas na questão da gestão e acesso aos dados científicos. O relatório é complementado ainda por um glossário.

2 – DADOS CIENTÍFICOS E REPOSITÓRIOS DE DADOS

2.1 – Investigação e dados científicos: contextualização e evolução recente

O registo das observações, ensaios e experiências, ou seja, a produção de dados, é já há vários séculos uma das características essenciais da ciência moderna. A forma e o volume desses registos ou dados científicos foram naturalmente evoluindo, crescendo em dimensão e complexidade, de acordo com a própria evolução da investigação científica, dos seus objectos, metodologias e instrumentos. De igual modo, foram-se registando alterações nas formas de armazenar, preservar, aceder e partilhar os dados produzidos no âmbito da actividade científica.

Apesar desta já longa história, a verdade é que desde meados do século XX e, sobretudo, nas duas últimas décadas, se registou uma verdadeira revolução no volume, complexidade e importância dos dados na actividade científica. Impulsionada por novos métodos, instrumentos e ferramentas, e apoiada nos progressos alcançados na capacidade computacional e no armazenamento digital, a investigação científica é cada vez mais intensiva na recolha dos dados em quase todas as áreas do conhecimento. Num número crescente de disciplinas, como a genética, medicina, física ou meteorologia, a investigação científica moderna depende da disponibilidade de grandes volumes de dados, organizados em bases de dados, públicas ou privadas, assim como da capacidade de os recuperar, recombinar e processar. Este tipo de ciência já é designado como *data-intensive science*[4] compreendendo, segundo os seus proponentes, três actividades essenciais: a recolha, a curadoria e a análise de dados.

Os dados científicos que, nos termos da definição da OCDE, são *"registos factuais usados como fontes primárias na investigação científica, e que são geralmente aceites na comunidade científica como necessários para validar os resultados de investigação"*[5], podem assumir várias formas (texto, números, imagens fixas, imagens em movimento, etc.) e dimensões, desde registos de observações individuais ou ensaios de pequenos laboratórios que não ultrapassarão algumas centenas de kilobytes, até aos dados

[4] Hey, Tony; Tansley, Stewart; Tolle, Kristin, eds. - The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, Washington, Microsoft Research, 2009.

[5] Tradução da definição de "Research data" (pág. 13) - *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris, 2007. Disponível em: http://www.oecd.org/dataoecd/9/61/38500813.pdf [consultado em: 27 de Abril de 2010].

produzidos pelo *Large Hadron Collider (LHC)*[6] do CERN, que pode gerar várias dezenas de petabytes por dia.

Os dados científicos podem ser classificados a partir de várias perspectivas. Por exemplo, a *National Science Foundation (NSF)*[7] dos Estados Unidos apresenta uma categorização das origens dos dados: dados de observação (registos históricos, como medidas de precipitação, que não podem ser reproduzidos e necessitam de preservação permanente), dados computacionais (resultantes de simulações, que teoricamente podem ser reproduzidos se for preservada informação sobre o modelo e a sua execução) e dados experimentais (que também não são facilmente reproduzíveis).

Os dados científicos são produzidos ou utilizados no contexto de investigação científica. Podem ser, por exemplo, recolhidos ou criados para efeito de processamento científico, como os dados atmosféricos usados para previsão meteorológica, ou os dados recolhidos de sensores para monitorizar o estado de um edifício. Há dados obtidos como resultados do processamento automático de objectos (por exemplo, uma colecção de imagens processada para obter os respectivos histogramas de cor, que constituem assim novos dados). Há ainda dados que não são produzidos para investigação mas que acabam por ser objecto dela, como as contribuições que os utilizadores de uma rede social fazem na forma de textos, fotografias ou outros objectos e que acabam por ser utilizados para estudos sociológicos.

Para se constituírem como verdadeiramente úteis, os dados científicos devem possuir estrutura e organização. Os conjuntos de dados (*datasets*) são uma das unidades essenciais. Os conjuntos de dados são colecções de informações ou factos relacionados entre si e registados num formato comum. Por exemplo, os resultados de um estudo de opinião por entrevista numa investigação sociológica constituem um conjunto de dados, composto pelos registos individuais das entrevistas.

As condições em que os dados recolhidos ou produzidos numa investigação podem ser acedidos e reutilizados por outros investigadores, para além do contexto em que foram gerados, são questões importantes para a ciência nos dias de hoje. A forma como são cuidados (curadoria de dados) e as condições legais associadas ao seu acesso e partilha constituem os dois elementos determinantes do futuro dos diversos conjuntos de dados científicos.

---

[6] Mais informações sobre o *Large Hadron Collider* do CERN acessíveis em: http://public.web.cern.ch/public/en/LHC/LHC-en.html.

[7] Sítio da *National Science Foundation* acessível em: http://www.nsf.gov.

Em primeiro lugar é preciso garantir que os dados são registados, mantidos e preservados de forma adequada. Um dos primeiros requisitos é que os conjuntos de dados sejam acompanhados de informação que descreva a sua origem (tempo ou espaço, métodos e instrumentos de recolha), âmbito, autoria, propriedade e condições de reutilização, ou seja, de metadados. Em paralelo com a interoperabilidade tecnológica, a existência de metadados adequados e normalizados é um requisito essencial para o acesso e reutilização dos dados científicos.

A curadoria dos dados não se esgota obviamente na criação de metadados e compreende o conjunto das acções que garantem a autenticidade, integridade e acessibilidade dos dados científicos. Em especial, a curadoria dos dados envolve todas as actividades de preservação necessárias para garantir a possibilidade de voltarem a ser usados no futuro.

2.2 – Dados científicos: do acesso à partilha

Para além da curadoria, o outro requisito para a utilização futura dos dados é a não existência de limitações legais (relacionadas com propriedade intelectual e direitos de autor) que o impeçam. Os perigos e prejuízos provocados pelas limitações à reutilização dos dados, sendo um problema com origem e manifestações anteriores, ganharam especial visibilidade e expressão há pouco mais de uma década, a propósito do Projecto Genoma Humano (PGH). O PGH, iniciado em 1990, foi um esforço internacional, financiado em grande escala por fundos públicos, para produzir o mapeamento do genoma humano e a identificação de todos os nucleótidos que o compõem. A ameaça real de que o acesso livre aos dados resultantes do PGH poderia ficar impedido por um pedido de patente solicitado por uma empresa privada que havia participado no projecto, a *Celera Genomics*, levou a que em Maio de 2000 Jim Kent, então um estudante de doutoramento em Biologia na Universidade da Califórnia, Santa Cruz (UCSC), criasse rapidamente um programa, utilizando o algoritmo *GigAssembler*[8], que possibilitou que os dados resultantes do Projecto Genoma Humano fossem recolhidos e disponibilizados publicamente através do *UCSC Genome Browser*[9].

[8] Mais informações sobre o algoritmo *GigAssembler* disponíveis em: http://genome.cshlp.org/content/11/9/1541.full.pdf.

[9] Sítio do *UCSC Genome Browser* disponível em: http://genome.ucsc.edu/cgi-bin/hgTracks?org=human.

Nos últimos anos generalizaram-se o conceito e os movimentos em favor dos dados abertos, ou *Open Data*. Se o termo *Open Data* é ainda relativamente novo, o conceito e a prática são já relativamente antigas[10].

O movimento dos dados científicos abertos preconiza que determinados dados sejam disponibilizados publicamente de forma gratuita, sem restrições de *copyright*, patentes ou outros mecanismos de controlo. Neste sentido, assemelha-se a uma série de outros movimentos de "abertura", tais como o *Open Source* ou o *Open Access*, que, contudo, possuem dinâmicas e objectivos próprios.

Dado que a maioria da investigação realizada a nível mundial é financiada por entidades públicas, a preocupação com o acesso e utilização dos dados científicos não se confina exclusivamente à comunidade científica, tendo entrado igualmente na agenda política. Uma das primeiras manifestações desta realidade ocorreu nos Estados Unidos, em 1995, quando o *Global Change Data and Information System* (GCDIS) começou por colocar em discussão pública o princípio da partilha integral e aberta de dados científicos[11].

Em Janeiro de 2004 registou-se outra iniciativa de grande relevância, quando Ministros da Ciência e Tecnologia de países da Organização para a Cooperação Económica e Desenvolvimento (OCDE), reunidos em Paris, discutiram a necessidade da existência de directrizes internacionais sobre o acesso aos dados científicos. Nesse encontro, os governos das 30 nações da OCDE, conjuntamente com a China, Israel, Rússia e África do Sul, subscreveram uma declaração[12] em que reconheceram a importância do acesso aos dados científicos gerados com financiamento público. No seguimento desta reunião, em Fevereiro de 2006, o *OECD's Committee for Scientific and Techonological Policy* indigitou um grupo de peritos para analisar e desenvolver um conjunto de directrizes baseadas em princípios comuns para promover o acesso aos dados científicos digitais resultantes de financiamento público. Após um extenso processo de consulta junto de instituições de investigação e de decisores políticos dos países membros da OCDE com vista a um consenso, os princípios e directrizes propostos pelo grupo de trabalho foram sancionados

---

[10] *"Although the term open data is rather new, the concept is rather old. The International Geophysical Year of 1957-8 caused the setting up of several world data centres and - more importantly - set standards for descriptive metadata to be used for data exchange and utilisation."* - Entrada de Keith G. Jeffery no Blog de Peter Murray-Rust: http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=32 [Consultado em 26 de Abril 2010].

[11] *International programs for global change research and environmental monitoring crucially depend on the principle of full and open exchange (i.e., data and information are made available without restriction, on a non-discriminatory basis, for no more than the cost of reproduction and distribution..."* Extracto de *On the Full and Open Exchange of Scientific Data* [Em linha] - Committee on Geophysical and Environmental Data National Research Council, Washington, D.C., 1995. Disponível em: http://www.nap.edu/readingroom.php?book=exch&page=exch.html [Consultado em 28 de Abril 2010].

[12] *OECD Declaration on Access to Research Data From Public Funding*, Paris, 2004. Disponível em: http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html [Consultado em 27 de Abril 2010].

em Outubro desse ano pelo comité e posteriormente apensos a uma recomendação que foi aprovada pelo Conselho Geral da OCDE a 14 de Dezembro de 2006[13].

Outro acontecimento relevante foi a criação, em 24 de Maio de 2004, da *Open Knowledge Foundation* (OKF)[14], uma organização sem fins lucrativos criada com o intuito de promover o acesso a conteúdos e dados abertos (designados pela fundação como *open knowledge*). A OKF publicou uma *Open Knowledge Definition*[15] e tem vindo a promover diversos projectos, tais como o *Comprehensive Knowledge Archive Network*[16] e a iniciativa *Open Data Commons*. Esta iniciativa tem vindo a ser desenvolvida através do seu Conselho Consultivo[17] e visa a promoção de soluções jurídicas que facilitem a abertura dos dados científicos. Em Março de 2008 foi apresentada a primeira licença neste domínio: *Public Domain Dedication and License (PDDL)*[18].

Uma das iniciativas mais recentes ocorreu em Julho de 2009, quando Peter Murray-Rust, Cameron Neylon, Rufus Pollock e John Wilbanks gizaram o primeiro esboço dos *Panton Principles*. Os *Panton Principles* incluem uma definição de dados abertos[19], recomendam que o estatuto jurídico dos conjuntos de dados científicos seja explícito, afirmam que as licenças de conteúdo não são apropriadas para os dados resultantes de investigação, desencorajam a utilização de licenças que restrinjam os usos comerciais e incentivam de forma veemente a consagração dos dados de investigação ao domínio público.
Na última década, o debate em torno do acesso e reutilização dos dados científicos tem sido bastante intenso. A figura seguinte apresenta uma cronologia de alguns eventos e iniciativas com maior relevo neste domínio.

---

[13] *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris, 2007. Disponível em: http://www.oecd.org/dataoecd/9/61/38500813.pdf [consultado em: 27 de Abril de 2010].

[14] Sítio da *Open Knowledge Foundation* acessível em: http://www.okfn.org.

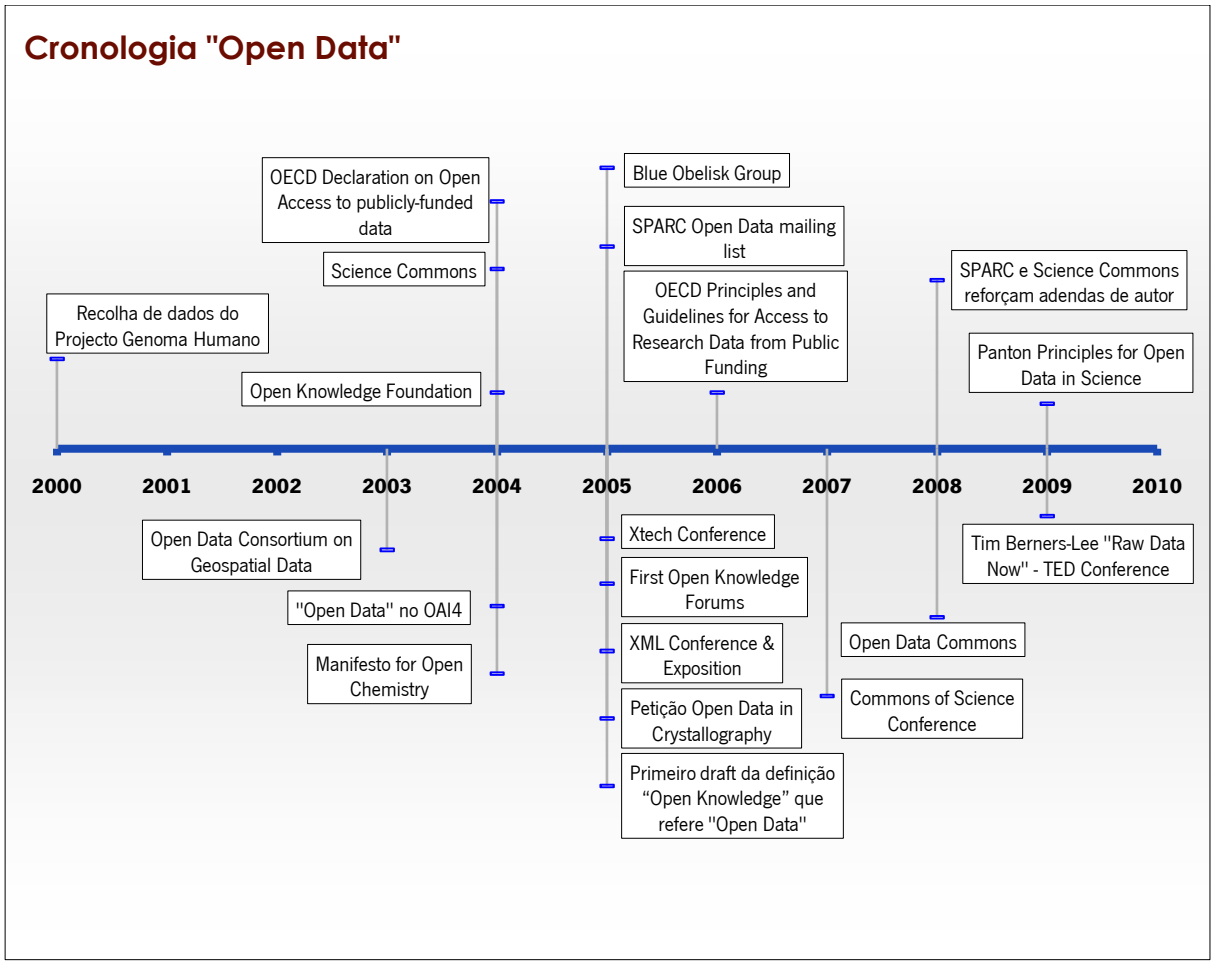
[15] *Open Knowledge Definition* disponível em: http://www.opendefinition.org/okd.

[16] Sítio da *Comprehensive Knowledge Archive Network* acessível em: http://www.ckan.net.

[17] Mais informações sobre o Conselho Consultivo da *Open Knowledge Foundation* acessíveis em: http://www.opendatacommons.org/about/advisory-council.

[18] Mais informações em sobre a licença *Public Domain Dedication and License (PDDL)* acessíveis em: http://www.opendatacommons.org/licenses/pddl/1-0/.

[19] *"By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. To this end data related to published science should be explicitly placed in the public domain"* – Extracto de *Panton Principles: Principles for Open Data in Science*. Disponível em: http://pantonprinciples.org [Consultado em 27 de Abril 2010].



2.3 – Curadoria de dados: as dimensões disciplinar e institucional

Como se pode concluir, regista-se a nível mundial um crescente reconhecimento dos méritos da partilha de dados científicos (*data sharing*), em princípio e na prática. Algumas instituições de financiamento da ciência começam a incorporar a partilha de dados como um dos seus requisitos. Há novas técnicas e ferramentas de software que facilitam a análise e novas explorações de dados. No entanto, verificam-se grandes diferenças no que diz respeito à curadoria e à partilha dos dados entre as várias áreas e disciplinas científicas, explicadas quer por factores culturais das diversas comunidades científicas quer pela existência, ou inexistência, das infra-estruturas adequadas (que vão desde os equipamentos até às normas de metadados). Algumas disciplinas, como a genética, as ciências climáticas e a astronomia, possuem infra-estruturas e normas bem implantadas que facilitam a pesquisa, acesso e reutilização dos dados científicos, enquanto noutras falta ainda quase tudo para que a partilha de dados se possa realizar de forma generalizada.

A necessidade da preservação de dados será também comum a todas as áreas disciplinares, mas o tipo de dados a preservar e o período de preservação é variável. As

estratégias de preservação dependem das características dos dados (volume, estrutura, qualidade, singularidade/originalidade, valor científico potencial para a disciplina) e a capacidade para de facto os preservar da existência de formatos de dados e esquemas de metadados apropriados, de infra-estruturas de armazenamento, capacidade técnica e evidentemente financiamento.

Foi recentemente publicado um relatório[20] sobre as diferenças entre as várias disciplinas no que diz respeito à curadoria e partilha de dados, com base no estudo comparativo de dezasseis casos. Considerando quatro grandes áreas disciplinares (artes e humanidades, ciências sociais, ciências da vida, ciências físicas), o relatório conclui que existem grandes diferenças não só entre elas, mas também dentro de cada uma delas.

Assim, nas artes e humanidades a partilha de dados é limitada, embora tenha alguma expressão em disciplinas específicas como a arqueologia, epigrafia e história da arte[21].

No ramo das ciências sociais, várias disciplinas recolhem e usam dados que possuem algumas limitações associadas a regras e acordos relativos a confidencialidade ou a considerações éticas ou legais. Esta realidade pode constituir uma barreira à partilha e reutilização de dados, mesmo considerando a possibilidade de os "obscurecer" (anonimizando ou descontextualizando)[22].

No domínio das ciências da vida, o volume de dados produzidos está a crescer dramaticamente. A dimensão dos *datasets* individuais pode ser muito grande, e a sua gestão e manipulação requerem a existência de grande capacidade de armazenamento e de computação. Apesar de, em teoria, existir uma ética de partilha de dados no domínio das ciências da vida, na prática essa partilha é limitada. Na área das ciências da saúde existem também várias limitações, constrangimentos e necessidades de salvaguardar dados.

Finalmente, no campo das ciências físicas existe uma grande variedade de práticas. Na astronomia a partilha de dados está bem estabelecida. Dentro das ciências climáticas existem variações significativas nas práticas de partilha de dados, mais comuns na modelação oceânica e nos dados observacionais e menos comuns, por razões comercias, na meteorologia e modelação climática. Outra área onde existem bons exemplos de partilha

[20] Key Perspectives. (2010), *"Data dimensions: disciplinary differences in research data sharing, reuse and long term viability: A comparative review based on sixteen case studies"*, DCC SCARP Synthesis Report comissioned by the Digital Curation Centre. Disponível em: http://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf [consultado em: 8 de Maio de 2010].
[21] Idem.
[22] Ibidem.

de dados é a cristalografia[23].

Para que os dados possam ser partilhados e reutilizados é necessário que existam organizações e indivíduos responsáveis pela sua curadoria. Os investigadores, por si só, não serão as pessoas adequadas para assegurar a preservação e o acesso continuado aos dados que recolhem e produzem. As suas competências concentram-se essencialmente no domínio da investigação, o que sugere que a responsabilidade pela curadoria de dados deve ser assegurada por outro tipo de profissionais. Estes profissionais podem possuir experiência disciplinar específica e serem integrados em grupos de investigação, departamentos ou unidades autónomas (e ser designados por *data scientists* ou *data managers*) ou podem ser peritos na área da informação que trabalhem em centros de dados ou em bibliotecas. Em qualquer caso, estes "novos" profissionais de dados necessitam de conhecimentos no domínio científico dos dados e formação específica sobre suporte e curadoria de dados[24].

Em algumas áreas existem centros de dados que podem assegurar uma curadoria profissional de dados considerados relevantes enquanto o volume de dados que processam não ultrapassar a sua capacidade. O problema é o que fazer com a "pequena ciência", isto é, com os conjuntos de dados produzidos por investigadores individualmente ou por pequenos grupos de investigação que não possuem nem os recursos nem as infra-estruturas para cuidar dos conjuntos de dados para além do termo dos seus projectos de investigação.

Dado o número crescente de repositórios institucionais, tem sido sugerido que eles possam ser a resposta, ou pelo menos parte dela, à necessidade de curadoria dos dados produzidos na investigação. De facto, existem já várias centenas de repositórios institucionais em estádio de produção, mas a sua utilização para albergar, preservar e dar acesso a conjuntos de dados científicos é ainda muito reduzida.

A curadoria dos dados produzidos pela comunidade científica da sua própria instituição constitui um desafio estratégico fundamental para os gestores e administradores dos repositórios institucionais. A existência da vontade, das competências e dos recursos necessários para fazer face a estes desafios no seio das instituições ainda necessita de ser completamente demonstrada.

---

[23] Ibidem.

[24] Swan, A. and Brown, S. (2008) The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Technical Report , School of Electronics & Computer Science, University of Southampton. Disponível em: http://eprints.ecs.soton.ac.uk/16675 [Consultado em 11 de Maio 2010].

Como se refere no estudo sobre as diferenças disciplinares que tem vindo a ser mencionado[25], para que os repositórios institucionais possam ser responsáveis pela curadoria de dados científicos das suas organizações terão de desenvolver estratégias específicas para cada disciplina, uma vez que uma abordagem genérica à curadoria de dados não será suficiente para lidar com todas as necessidades e expectativas dos investigadores das diferentes áreas. Ora, é precisamente esta necessidade de conjugar a dimensão institucional (muito ampla e multidisciplinar no caso das universidades) com a dimensão disciplinar (com os seus requisitos específicos) que constitui um dos principais desafios à utilização dos repositórios institucionais como uma componente fundamental na infra-estrutura global de curadoria dos dados científicos.

2.4 – Repositórios de dados científicos: situação actual

De uma forma geral, os conjuntos de dados científicos constituem, ao nível da maior parte das organizações que produzem ou lidam com dados em grande escala, uma preocupação. Isto significa que há alguma sensibilidade dos cientistas e administradores à fraca robustez das actuais infra-estruturas, mas ainda não se chegou ao ponto de a curadoria destes dados ser um problema reconhecido e contemplado em projectos, políticas das organizações, orçamentos ou constituição de equipas técnicas.

Há, no entanto, uma grande diversidade de iniciativas como se pode constatar em diferentes directórios de repositórios. É o caso do OAD[26] (Open Access Directory, um wiki do Simmons College, Boston - EUA), que lista repositórios de dados e bases de dados por domínios científicos. A Google está também a lançar, no âmbito dos "Google Labs", o "Google Public Data Explorer"[27], um wiki para repositórios de dados que fornece ferramentas para a sua exploração.

O estado de desenvolvimento das infra-estruturas de suporte aos dados é muito diverso. Para poder comparar abordagens que tenham pressupostos semelhantes, apresentam-se a seguir 5 cenários que correspondem a formas diferentes de tratar a curadoria dos dados.

[25] Ibidem.

[26] Sítio da *Open Access Directory* acessível em: http://oad.simmons.edu/oadwiki/Data_repositories.

[27] Sítio do *Google Public Data Explorer* acessível em: http://www.google.com/publicdata/directory.

1. Curadoria pelos cientistas ou técnicos que usam os dados

Este é um cenário em que não há uma política institucional para a curadoria e em que esta decorre da preocupação dos agentes envolvidos. A eficácia desta gestão depende de circunstâncias particulares, tais como os dados serem usados de forma sistemática ou ser missão da instituição fornecê-los a terceiros.

Este cenário encontra-se facilmente em qualquer universidade ou centro de investigação, em particular em grupos cujas actividades têm uma componente importante de processamento de dados e em áreas em que não existem ainda requisitos de registo de dados em formatos normalizados, nem uma prática estabelecida de troca de dados entre organizações. Neste cenário, há alguma garantia de que os dados se mantenham activos, embora a generalização do uso de máquinas pessoais poderosas possa constituir um risco. Verifica-se com frequência que, mesmo em instituições com infra-estruturas informáticas sólidas, os dados em máquinas individuais estão frequentemente sem salvaguarda sistemática.

Repositórios como o CAVA[28], "Human Communication: an Audiovisual Archive", uma iniciativa do UCL (University College, Londres) e o OASIS[29], "Open Access Series of Imaging Studies", que disponibiliza conjuntos de dados de ressonâncias magnéticas cerebrais, patrocinado por um conjunto de universidades e institutos de investigação, estão no limite deste cenário. Estes repositórios foram criados através de financiamentos de investigação e são mantidos com os meios que as respectivas agências fornecem.

2. Curadoria por organizações científicas sectoriais, instituições geradoras, recolectoras ou distribuidoras de dados

Este cenário surge quando é identificada a necessidade de realizar acções de reunião e de preservação de conjuntos de dados, ou de fornecer serviços de acesso dentro de uma comunidade ou de um domínio de investigação. Tipicamente estas iniciativas envolvem esforço voluntário de universidades ou de outras instituições e beneficiam do apoio dos organismos financiadores da investigação, que começam a reconhecer a sua importância. Este cenário encontra-se em muitas associações científicas que instalam ou contratam infra-estruturas de suporte aos seus recursos de dados. O acesso a estes dados pode ser restrito aos cientistas envolvidos ou ter um âmbito mais lato de divulgação pública. O cenário tem como ponto forte o facto de reunir comunidades com interesse sustentado nos dados e de

[28] Mais informações sobre a iniciativa *Human Communication: an Audiovisual Archive* (CAVA) disponíveis em: http://www.jisc.ac.uk/whatwedo/programmes/inf11/sue2/cava.aspx.

[29] Sítio do *Open Access Series of Imaging Studies* (OASIS) acessível em: http://www.oasis-brains.org/.

ter origem nos próprios investigadores, que podem fornecer descrições especializadas dos seus dados. A capacidade destas estruturas de suporte de dados garantirem o acesso aos dados que albergam, mesmo quando estes deixem de ser usados activamente, é variável.

Um exemplo deste cenário na área das artes, humanidades e ciências sociais é o DANS – "Data Archiving and Networked Services"[30], na Holanda, lançado pela KNAW (Royal Netherlands Academy of Arts and Sciences) e suportado pelo NWO (Netherlands Organisation for Scientific Research). Este repositório fornece serviços como o depósito de conjuntos de dados, a descarga de dados, ferramentas para a análise de dados e ligação a dados alojados em outros repositórios.

Um outro exemplo, na área da biotecnologia, é o NCBI[31] (National Center for Biotechnology Information nos EUA) que reúne recursos nas áreas da biomedicina e do genoma. Neste caso, o suporte a dados é de natureza mais operacional do que descritiva: o centro mantém um conjunto de bases de dados especializadas e directamente utilizadas pelos investigadores. O uso destes dados supõe familiaridade com as suas representações.

3. Curadoria por universidades ou centros de investigação

Este cenário é semelhante ao anterior, mas a iniciativa parte de uma universidade ou centro de investigação e por isso tende a incluir dados de áreas diversas. Também aqui é usado o apoio dos organismos financiadores. Neste cenário podem ser diversas as formas de integrar a curadoria dos dados na instituição, sendo duas das mais frequentes a utilização dos centros de computação e as bibliotecas. Os centros de computação encontram-se bem apetrechados no que se refere a serviços e pessoal com perfil informático, mas não têm habitualmente recursos com formação apropriada para apoiar a modelação de dados e a descrição dos conjuntos de dados. As bibliotecas podem oferecer recursos para descrição, mas não a descrição especializada requerida pelos conjuntos de dados, e não costumam ter recursos informáticos próprios.

Nas universidades e centros de investigação a divulgação de dados científicos tem como objectivo adicional a visibilidade pública da instituição. Um ponto forte deste cenário é o facto de tirar partido das infra-estruturas informáticas da organização, o que torna os serviços oferecidos mais robustos. Por outro lado, tendo de acomodar conjuntos de dados

[30] Sítio do *Data Archiving and Networked Services* (DANS) acessível em: http://www.dans.knaw.nl/.

[31] Sítio do *National Center for Biotechnology Information* (NCBI) acessível em: http://www.ncbi.nlm.nih.gov/.

de áreas muito diversas, tenderá a ter uma ligação menos estreita aos produtores dos dados e menor especialização na descrição.

As bibliotecas associadas a instituições de investigação estão neste momento muito motivadas para a criação e gestão de repositórios institucionais com recolha sistemática da produção interna. Os repositórios de dados têm aparentemente uma ligação forte a estas iniciativas: por um lado, também requerem infra-estruturas com requisitos de registo, descrição, pesquisa e acesso; por outro, começa a ser frequente as publicações científicas terem associados dados científicos. Para uma biblioteca poder assegurar a curadoria dos dados, é necessário que detenha os recursos técnicos que lhe permita trabalhar de forma próxima com os investigadores que os produzem. Um ponto forte deste cenário é o facto de criar uma infra-estrutura durável. Possíveis fragilidades são a distância que se estabelece entre o serviço e os investigadores e criadores de conjuntos de dados e a necessidade de acomodar dados de natureza muito diversa, com prejuízo da descrição especializada.

O repositório Datashare[32], na Universidade de Edimburgo, Escócia, é um exemplo de serviço avançado de partilha de dados no âmbito de uma biblioteca. Este cenário é pouco usual, e decorre do facto de a universidade ter, sob a designação de "Information Services", reunido o EDINA[33], um centro de dados académicos suportado pelo JISC, e a "University Data Library".

4. Curadoria por organismos oficiais

Neste cenário encontram-se as iniciativas que partem dos organismos de gestão da ciência ao nível de um país. Existem normalmente condições prévias que criam a sensibilidade ao problema e são criadas infra-estruturas de suporte mantidas ao nível de organismos nacionais.

Existem várias instâncias do cenário 4 nos países onde a sensibilidade à curadoria dos dados científicos é maior. Um ponto forte deste cenário é o facto de criar uma infra-estrutura durável. Possíveis fragilidades são a distância que estabelece entre o serviço e os investigadores e criadores de conjuntos de dados e a necessidade de acomodar dados de natureza muito diversa, com prejuízo da descrição especializada.

[32] Mais informações sobre o repositório Datashare disponíveis em: http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf.

[33] Sítio do *EDINA* acessível em: http://edina.ac.uk/.

Um exemplo deste cenário é o ANDS[34]- "Australian National Data Service", financiado por diversos organismos oficiais e que tem por objectivo favorecer a visibilidade na web de dados científicos australianos, promover a curadoria dos dados e contribuir para as políticas da sua gestão. A NBII[35] (National Biological Information Infrastructure, EUA) é outro exemplo, na forma de um portal de acesso à informação de agências governamentais, académicas, organizações não governamentais e indústria, suportado por parcerias com as organizações produtoras.

Como exemplo de uma iniciativa recente, o DataONE[36] (Data Observation Network for Earth) é uma fundação, financiada pela NSF (National Science Foundation, EUA) e estabelecida em Fevereiro de 2010 para suportar a investigação e a divulgação de dados relacionados com as ciências do ambiente.

5. Curadoria por comunidades informais em linha

Este é um cenário que possivelmente se vulgarizará nos próximos anos, com o aparecimento de comunidades que reúnem especialistas e amadores na recolha e descrição de conjuntos de dados. O ponto forte deste cenário é a sua "robustez informal": com comunidades dispersas, é provável que existam múltiplas réplicas de um conjunto de dados, o que previne o seu desaparecimento. Por outro lado, não há um modelo institucional que garanta o acesso aos dados e pode haver dificuldade em identificar os responsáveis pela sua criação.

Exemplos deste cenário podem encontrar-se em colecções de dados como os do Wikispecies[37], um wiki que recolhe informação taxonómica e de descrição de espécies, com contribuições do público. Um conjunto de dados como este pode ser usado tanto para trabalho científico como para a divulgação científica.

2.5 – Tecnologias e plataformas dos repositórios de dados

No presente contexto o termo “repositório” designa um sistema informático em que existe uma plataforma de armazenamento de objectos representados em ficheiros, capaz de incorporar novos objectos à medida que são produzidos ou submetidos. O repositório oferece serviços que são dirigidos a quem deposita, a quem pesquisa e aos administradores

[34] Sítio do projecto *Australian National Data Service* (ANDS) acessível em: http://ands.org.au/.

[35] Sítio da *National Biological Information Infrastructure* (NBII) acessível em: http://www.nbii.gov/.

[36] Sítio da iniciativa *Data Observation Network for Earth* (DataONE) acessível em: https://dataone.org/.

[37] Sítio do *Wikispecies* acessível em: http://species.wikimedia.org/wiki/Main_Page.

do sistema. Nos repositórios de dados pode ir-se muito além desta visão de repositório de objectos, uma vez que cada conjunto de dados tem características próprias e por isso pode requerer um tratamento diferenciado. Como exemplo de um processamento específico de um conjunto de dados, pode referir-se a transformação de dados originalmente organizados numa base de dados para um formato textual (por exemplo na linguagem de anotação XML). Por outro lado, para o utilizador do repositório pode ser importante, por exemplo, em vez de descarregar a totalidade de um conjunto de dados cujo volume seja difícil de gerir, seleccionar apenas um subconjunto dos itens incluídos no conjunto, ou apenas alguns dos atributos desses itens.

Considerando o que pode conter um conjunto de dados científicos, as acções de curadoria podem levar a actuar a três níveis, que aqui se designam de armazenamento, de representação do conjunto e de representação do item. A designação “conjunto de dados” supõe que neste se podem identificar elementos individuais, ou itens. O primeiro nível, do armazenamento, trata de garantir que o conjunto de dados está guardado de forma fiável, num sistema com garantia de manutenção e salvaguarda. O nível 2, da representação do conjunto, assegura que existe descrição do conjunto de dados, garantindo metadados tais como os que dizem respeito à sua produção, direitos de uso ou características técnicas. O nível 3, da representação do item, lida com as questões de modelo e descrição para itens individuais e é dependente da natureza do conjunto de dados. Num conjunto de fotografias, a descrição ao nível do item pode incluir características técnicas da fotografia, do evento retratado e do fotógrafo. Num conjunto de dados recolhidos por sensores numa estrutura de betão, a descrição pode enumerar as grandezas recolhidas, as características dos sensores e os pormenores do ambiente em que a experiência foi montada. Num conjunto de dados sobre estruturas químicas, a descrição pode incluir referências aos elementos presentes e às suas ligações, bem como a informação necessária para os representar em 3 dimensões.

Nos repositórios de dados actuais são utilizadas diversas soluções tecnológicas. Algumas destas soluções são plataformas desenvolvidas genericamente para repositórios, outras são soluções desenvolvidas à medida para um caso específico. Esta divisão está relacionada tanto com os modelos de curadoria, identificados na Secção 2.3, como com o tipo de uso a que os conjuntos de dados estão sujeitos. Os modelos de curadoria podem ser vistos como um fluxo de operações, envolvendo uma equipa para além dos próprios investigadores, com a missão de fazer o acompanhamento e a disponibilização dos dados. Quando os responsáveis por esta equipa têm de a desenhar para apoiar diversas áreas disciplinares, a sua infra-estrutura de suporte tende a ser uma plataforma mais genérica, que desempenhe as funções básicas de curadoria sem se especializar em algum tipo de dados. Quando, pelo

contrário, a equipa é da responsabilidade directa dos cientistas de uma área disciplinar, ela tende a ser orientada para funções mais específicas de apoio à investigação.

No presente, podem-se identificar dois tipos de práticas nas infra-estruturas de suporte a repositórios de dados. A primeira existe em domínios que têm já uma prática estabelecida de registo e partilha de dados, seja porque as regras para publicação o requerem, seja porque os grupos de investigação não têm meios para recolher dados próprios e a investigação é baseada em dados recolhidos por organismos supervisores. Estão neste caso os repositórios de dados de genoma e os de dados astronómicos. Nestes domínios pode-se dizer que o paradigma vigente é a base de dados, e não o repositório de dados. Isto significa que, dos três níveis de curadoria anteriormente identificados, estão assegurados o 1 (armazenamento) e o 3 (representação do item). As bases de dados são instaladas em estruturas tecnológicas próprias, pelo que o nível de armazenamento está garantido. Como os dados são usados por diversas equipas, é necessário um conhecimento preciso e completo da forma como eles estão representados, pelo que a descrição ao nível do item é cuidada, embora possa estar escondida em modelos de dados e interfaces de acesso que apenas serão acessíveis aos investigadores do meio. Esta infra-estrutura de suporte em base de dados está tipicamente associada a um modelo de curadoria muito próxima dos investigadores, o cenário 2 mencionado na Secção 2.4.

A segunda prática está a surgir actualmente, impulsionada em grande medida pelo movimento do acesso livre. Entidades oficiais de financiamento, universidades, centros de investigação e agências governamentais estão a tomar iniciativas ou a ser pressionadas no sentido de exporem os dados recolhidos ou usados em projectos financiados. Como estas acções são frequentemente entregues aos mesmos serviços que gerem os repositórios institucionais, e os próprios conjuntos de dados estão muitas vezes ligados às publicações que neles se baseiam, o paradigma aqui é o de uso das plataformas de repositórios, das quais as mais divulgadas são as de desenvolvimento com licenças de código aberto, que surgiram para dar resposta à necessidade de coleccionar e preservar a produção de natureza científica e técnica dentro de comunidades de investigação. O seu uso para os repositórios de dados dá ênfase aos níveis de curadoria 1 e 2, uma vez que são capazes de lidar com o armazenamento e a descrição ao nível da colecção, mas não com a descrição ao nível do item. Esta infra-estrutura de suporte em plataformas de repositórios está normalmente associada a modelos de curadoria por universidades ou por organismos oficiais, os cenários 3 e 4 mencionados na Secção 2.4.

2.5.1 – Plataformas

Estão hoje disponíveis diversas plataformas de repositórios, que foram desenvolvidas para recolher, preservar e divulgar literatura científica, mas presentemente podem ser usadas para agregar outros tipos de conteúdos digitais.

DSpace

É a ferramenta de repositórios mais divulgada e tem uma extensa lista de utilizadores. Foi lançada pelas MIT Libraries e pelos Hewlett-Packard Labs em 2002 com o objectivo de fornecer um sistema de repositório para documentos digitais resultantes de investigação ou destinados à educação e distribuído com uma licença de código aberto. A partir de 2007, o MIT e a HP criaram a Dspace Foundation[38], uma organização sem fins lucrativos para promover a plataforma e suportar os seus utilizadores. Em 2009, este suporte passou para a DuraSpace Foundation, também uma organização sem fins lucrativos dedicada a "tecnologias de código aberto e da nuvem para biblioteas, universidades, centros de investigação e organizações do património cultural". A lista de instâncias da plataforma DSpace presente no sítio web tem mais de 800 entradas[39].

Exemplos de repositórios de dados que usam a plataforma DSpace são o Edinburgh DataShare[40] da Universidade de Edimburgo, que é especializado em conjuntos de dados. Os repositórios institucionais do MIT e da Universidade de Cambridge são dois exemplos entre muitos dos que aceitam conjuntos de dados como um dos tipos de objectos depositados, organizando-os no repositório geral.

EPrints

O "EPrints Repository Software"[41] é também uma plataforma muito divulgada para repositórios institucionais. Foi criada e é mantida pela "School of Electronics and Computer Science" da Universidade of Southampton, Reino Unido. A plataforma é distribuída com base numa licença de código aberto. Além de oferecer as funcionalidades comuns nos repositórios institucionais, tem associado o "EPrints Services", com uma equipa de consultoria que pode acompanhar um projecto de instalação de um repositório desde a análise e desenvolvimento personalizado até ao fornecimento do serviço de gestão. Um dos desenvolvimentos recentes é a integração no repositório de ferramentas que estão a surgir

[38] Sítio da fundação DuraSpace acessível em: http://duraspace.org/.

[39] Mais informações disponíveis em: http://www.dspace.org/whos-using-dspace/Repository-List.html.

[40] Sítio do repositório *DataShare* acessível em: http://datashare.is.ed.ac.uk/dspace/.

[41] Sítio do *EPrints Repository Software* acessível em: http://www.eprints.org/.

na comunidade de preservação digital. Encontra-se no sítio web uma lista de mais de 250 instâncias de repositórios[42].

Um exemplo de repositório de dados baseado em EPrints é o repositório de dados eCrystals - Southampton[43], um arquivo de estruturas de cristais gerado pelo grupo de Cristalografia Química da Universidade de Southampton e pelo "EPSRC UK National Crystallography Service".

Fedora

O Fedora não é uma plataforma para repositórios como o Dspace ou o EPrints, mas uma arquitectura extensível que pode ser usada para desenvolver software para repositórios. Criada pela Universidade de Cornell, é actualmente mantida pela fundação DuraSpace, tal como o DSpace. Tem princípios como o da agregação de conteúdos locais e distribuídos em objectos digitais e a associação destes a serviços [Fedora]. A arquitectura inclui ainda um modelo de relações baseado no RDF (Resource Description Framework)[44] do W3C usado para ligar os objectos aos seus componentes. Está disponível em licença de código aberto e tem sido usado em diversas aplicações para bibliotecas digitais, arquivos, repositórios institucionais e sistemas de objectos de aprendizagem. A lista de utilizadores do software Fedora presente no sítio web tem mais de 100 entradas, incluindo um grande número de bibliotecas de destaque: a Biblioteca do Congresso dos EUA, as bibliotecas nacionais de Portugal, de França, da Austrália e a Biblioteca Pública de Nova Iorque são exemplos de desenvolvimentos baseados nesta plataforma[45].

Um exemplo de repositório de dados baseado em Fedora é o DataBank da Universidade de Oxford[46]. Outros repositórios institucionais que usam Fedora incluem conjuntos de dados a par de publicações. Um exemplo é o "Center for International Earth Science Information Network", da Universidade de Columbia, com dados de áreas muito diversas[47].

[42] Mais informações disponíveis em: http://www.eprints.org/software/archives/.

[43] Sítio do repositório *eCrystals - Southampton* acessível em: http://ecrystals.chem.soton.ac.uk/.

[44] Mais informações disponíveis em: http://www.w3.org/RDF/.

[45] Mais informações disponíveis em: https://fedora-commons.org/confluence/display/FCCommReg/Fedora+Commons+Registry.

[46] Sítio do repositório *DataBank* acessível em: http://databank.ouls.ox.ac.uk/.

[47] Mais informações disponíveis em: http://www.ciesin.org/.

eSciDoc

O eSciDoc[48] é uma plataforma desenvolvida por iniciativa da Max Planck Society e do FIZ Karlsruhe, financiada pelos organismos oficiais alemães, e que tem como objectivo suportar organizações de investigação multi-disciplinares. Esta plataforma pretende ser um ambiente de suporte à chamada *e-research*, oferecendo serviços para apoiar a colaboração. Relativamente à função repositório, pretende incluir tanto as publicações como a documentação intermédia, os dados e os materiais de aprendizagem[49]. No que concerne ao software de repositório, o eSciDoc é baseado em Fedora. No seu estado actual, a plataforma apresenta como exemplos de casos de utilização desenvolvidos os repositórios de publicações, um ambiente de sala virtual para explorar colecções digitais e o "Scholarly Workbench", um ambiente colaborativo para comunidades nas artes e humanidades. O DANS – "Data Archiving and Networked Services", na Holanda, é baseado no eSciDoc.

2.5.2 – Metadados

Uma questão que aparenta estar ainda pouco desenvolvida nas diversas iniciativas relativas aos repositórios de dados científicos é a normalização dos metadados. Esta situação não é de estranhar, atendendo a que, por um lado, falta ainda resolver muitos dos problemas básicos, e por outro é de prever que seja muito difícil uniformizar a descrição entre domínios científicos.

A descrição dos conjuntos de dados ainda se pode considerar muito pouco desenvolvida. Ao nível do conjunto, têm sido adoptados modelos genéricos como o do Dublin Core, já muito utilizado nos repositórios institucionais. Ao nível do item, não se encontram ainda experiências reportadas. Em alguns casos como o do projecto DataShare, que envolveu equipas de Edimburgo, Oxford e Southampton, os resultado incluem algumas recomendações quanto a extensões a introduzir nos metadados usados nos repositórios institucionais[50].

Uma iniciativa mais centrada nos dados científicos está actualmente a ser lançada na comunidade Dublin Core. Trata-se do "DCMI Science and Metadata Community"[51], que é um forum para a troca de informação sobre metadados para descrever dados científicos, dirigida aos responsáveis pela curadoria de dados. Esta comunidade centra o seu trabalho

[48] Matthias Razum, Frank Schwichtenberg, Steffen Wagner, Michael Hoppe. 2009. eSciDoc Infrastructure: A Fedora-Based e-Research Framework. M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 227-238.

[49] Mais informações disponíveis em: https://www.escidoc.org/JSPWiki/en/Overview.

[50] Mais informações disponíveis em: http://www.disc-uk.org/datashare.html.

[51] Mais informações disponíveis em: http://dublincore.org/groups/sam/.

nos desafios colocados pela curadoria dos dados e nas soluções que podem ser baseadas na arquitectura e princípios da Dublin Core Metadata Initiative.

Relativamente aos repositórios de dados pode concluir-se que há muito a fazer tanto no estabelecimento de práticas de descrição adequada como na normalização dos descritores a usar. A descrição normalizada dificilmente poderá cobrir toda a especificidade de um conjunto de dados, mas muitos dados descritivos a nível de conjunto ganharão em ser normalizados.

3 – REPOSITÓRIOS, PARTILHA E CURADORIA DE DADOS

3.1 – Curadoria e partilha de dados: problemas e desafios

A tomada de consciência da necessidade do armazenamento e da preservação de dados científicos em repositórios criados e mantidos para esse efeito constitui um processo ainda em curso, com diferentes estádios de maturidade a nível internacional. Por norma, é aos investigadores que se atribui a responsabilidade pela não partilha dos dados resultantes de processos de investigação - quer porque não estarão suficientemente sensibilizados para aspectos de preservação e de acesso à produção científica, quer porque poderão não estar interessados ou sentir relutância em disponibilizar, mesmo não sendo em acesso aberto, os dados que recolheram, quer, ainda, pelo facto de poderem não dispor de competências técnicas específicas para o efeito. A realidade, porém, é mais complexa, pois os investigadores não trabalham isolados, mesmo quando é frágil o seu vínculo institucional.

As entidades cuja missão inclui a realização e promoção da investigação científica são de igual modo responsáveis pela recolha, preservação e garantia do acesso aos dados produzidos. O que sucede com frequência é que, à semelhança do que se passa com os investigadores, estas instituições apenas raramente estão despertas para a necessidade de preservar esses dados, recolhendo-os, zelando pela sua manutenção e colocando-os à disposição de quem deles necessita. Da mesma forma, só a título excepcional dispõem de recursos e de competências para criar e assegurar a sustentabilidade de repositórios cuja função primordial consista na curadoria de dados científicos, recolhidos e produzidos durante actividades de investigação. Neste contexto, afigura-se como indispensável a aproximação entre os investigadores e as instituições que gerem repositórios para alojamento, preservação e acesso a dados científicos.

3.1.1 – Necessidades e expectativas dos investigadores

As políticas, as estratégias e as práticas de gestão de informação não poderão assegurar a curadoria e o acesso aos dados científicos sem que se disponha de conhecimentos fiáveis acerca da forma como se processa a investigação nas áreas em que se pretenda intervir. Por exemplo, é conveniente que os gestores de repositórios conheçam como é que os investigadores organizam os dados recolhidos e como os transformam em resultados. Do mesmo modo, o conhecimento do *workflow* dos dados poderá ser um contributo da maior importância para que se identifiquem os serviços de suporte necessários aos investigadores, tendo em vista o depósito, a preservação, o acesso e, sempre que possível, a divulgação

dos dados recolhidos. Como exemplo de boas práticas, registe-se que, no Reino Unido, os investigadores são obrigados a tornar disponíveis os dados da investigação, mas também a informar o que planeiam fazer para que possam ser partilhados. Para este efeito, os investigadores procedem à caracterização das bases de dados que pretendem criar, indicam as normas que irão seguir e o modo como planeiam disponibilizar os dados recolhidos, assim como identificam a entidade que se responsabilizará pela sua gestão global, como, por exemplo, pela preservação a longo prazo[52].

As necessidades dos investigadores começam a montante da investigação propriamente dita, isto é, no momento em que a projectam. Pode mesmo afirmar-se que o apoio de que necessitam se situa logo aí, quando indagam se já existirão dados recolhidos sobre o seu objecto de estudo e, em caso afirmativo, se os podem utilizar. A esta necessidade soma-se a da utilização de recursos, serviços e ferramentas e, por consequência, a de receber a formação necessária para o respectivo uso.

Noutro plano situam-se as dificuldades sentidas em matéria de protecção da confidencialidade dos dados recolhidos e tratados, nomeadamente quando depositados e disponibilizados em repositórios digitais. Por fim, é natural que os investigadores necessitem de instruções e de apoio técnico para poderem disponibilizar, partilhar e disseminar os dados recolhidos e tratados.

Sintetizando, poder-se-á afirmar que as necessidades essenciais dos investigadores se traduzem nos seguintes itens:

- Solução segura e amigável para armazenamento de grandes volumes de dados e para a sua partilha controlada;
- Infra-estrutura sustentável para publicação e preservação a longo termo de dados de investigação;
- Instruções para questões práticas relacionadas com a gestão dos dados ao longo do seu ciclo de vida;
- Orientações no que se refere à publicação de dados e à sua preservação;
- Apoio (competências, ferramentas e normas) durante todo o ciclo de vida da investigação, tendo em vista o armazenamento, a preservação e a garantia de que os dados recolhidos poderão vir a ser acedidos e reutilizados[53].

[52] Digital Repository Services for Managing Research Data: What Do Oxford Researchers Need? Luis Martinez-Uribe, Digital Repositories, Research Co-ordinator, Oxford e-Research Centre, University of Oxford. [*c.* Dez. 2008].
[53] Idem.

3.1.2 – Benefícios da partilha de dados

Deixando de lado aspectos de natureza ética e legal, impeditivos do acesso a dados científicos, parece legítimo concluir que a adesão dos investigadores a uma política generalizada de partilha de dados depende, antes de mais, da assunção dos benefícios individuais e colectivos que dela resultam, embora também de uma real tomada de consciência dessas vantagens por parte das instituições que os enquadram do ponto de vista institucional e académico e suportam financeiramente os custos da investigação.

Entre as diversas vantagens decorrentes do acesso aos dados de investigação e à sua partilha constam as oportunidades acrescidas de uso e de reutilização de dados já recolhidos, evitando-se, desta forma, investigações redundantes ou repetidas e os custos a elas inerentes. Na tentativa de estimular a inovação e evitar duplicações, são numerosas as entidades financiadoras de actividades de investigação que, em todo o mundo, têm produzido orientações e sugestões e divulgado mandatos para que se tornem públicos os dados produzidos.

Quando reutilizados, os dados poderão converter-se em embriões de novos projectos, assim como conduzir à experimentação e à verificação de hipóteses não consideradas aquando da investigação original. A reutilização de dados científicos poderá, também, dar lugar a novas áreas de trabalho, potenciadoras de empregabilidade, de incremento do investimento público e de produção de riqueza. E, embora o retorno do financiamento varie em função dos diferentes domínios de investigação, não restam dúvidas de que a não duplicação de recolha de dados constitui, por si só, um benefício.

Analisada sob outro prisma, a possibilidade de acesso a dados brutos e a informação primária resultante de investigação poderá constituir um valioso contributo em matéria de educação da comunidade de investigadores e de formação das gerações vindouras. Por seu turno, a validação de métodos de estudo e de análises técnicas, no que se refere à detecção de inexactidões, por exemplo, só é possível mediante a utilização de dados brutos.

Embora não constituindo condição imprescindível à transparência da investigação praticada, o facto de os dados criados poderem ser acedidos por outrem contribui para assegurar essa transparência. Ainda que se argumente, por vezes, que a acessibilidade aos dados de investigação constitui um incentivo à prática de fraudes, tem sido demonstrado o oposto, isto é, que a publicação de trabalhos científicos em acesso aberto, assim como a divulgação dos resultados da investigação que lhes é subjacente, contribuem para a detecção de fraudes, para a dissuasão de plágios e, ainda, para um registo mais completo e transparente da

ciência. Recorde-se, neste contexto, situações de fraude científica já ocorridas, que conduziram investigadores e editoras a facultar os meios de prova dos resultados de investigação obtidos, antes, ainda, da respectiva publicação[54].

Tanto para a ciência em geral, como para os investigadores, poderá ser muito benéfica a publicação dos resultados da investigação, de preferência em universos amplos, susceptíveis de múltiplos acessos. Uns porque adquirem a visibilidade de que não dispunham, outros porque assistem ao reforço do reconhecimento de que já usufruíam. Por seu turno, a avaliação e a revisão feitas pelos pares e as próprias candidaturas a financiamentos podem transformar-se em processos desburocratizados quando as partes intervenientes dispõem de acesso à produção dos investigadores e ao seu impacto no progresso científico. O reconhecimento dos investigadores traz consigo o das respectivas comunidades ou das instituições que suportam financeiramente a investigação, assim como o dos repositórios que alojam, preservam e disseminam os dados produzidos. A estes benefícios acresce, ainda, o impacto na qualidade e na eficiência da investigação, embora este possa não ser visível e mensurável no imediato.

Registe-se, também, como um benefício decorrente da partilha de dados, o acréscimo de oportunidades para as "indústrias" de reutilização de dados, como as de investigação geo-espacial, meteorológica e oceanográfica[55]. Mais oportunidades haverá, também, para o aparecimento de novas "indústrias", fornecedoras de produtos e serviços dirigidos à simplificação de armazenamento e de acesso a bases de dados.

3.1.3 – Inibidores da partilha de dados

Apesar do atrás exposto, a cultura da reutilização de dados não se encontra, ainda, enraizada na comunidade de investigadores. Os benefícios explícitos da disponibilização e da partilha de dados científicos são ainda muito escassos ou inexistentes. Atendendo ao facto de que os dados obtidos no decurso de processos de investigação constituem uma componente crítica essencial do capital intelectual de quem investiga e que o trabalho desenvolvido pelos investigadores exige muito esforço e se processa à custa de conhecimentos especializados, assumindo, com frequência, o estatuto de investimento de longa duração, torna-se compreensível que a concessão de acesso aos dados brutos seja muito rara.

[54] Adrian Burton, Andrew Treloar - Publish My Data: A composition of services from ANDS and ARCS. Australian National Data Service (ANDS). {Canberra|Melbourne}, Australia.

[55] Cf. Jenny Fry, Suzanne Lockyer and Charles Oppenheim - "Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes". Department of Information Science, Loughborough University; John Houghton and Bruce Rasmussen, Centre for Strategic Economic Studies, Victoria University, Melbourne. Novembro de 2008.

Acresce, ainda, o receio de que os dados produzidos sejam explorados e reutilizados de forma incorrecta por outros investigadores, para além de que há quem não pretenda reutilizar dados recolhidos no âmbito de outros processos de investigação, atendendo às diferenças que existem na concepção de processos experimentais e na recolha de dados.

Aos aspectos referidos somam-se, ainda, a preocupação com a confidencialidade de certo tipo de dados, pois nem sempre é possível a publicação, em acesso aberto, de dados recolhidos em determinadas circunstâncias, para não mencionar os períodos de embargo a que muitos deles se encontram sujeitos. Neste âmbito, enquadram-se, também, constrangimentos de natureza ética, que restringem a publicação de dados em repositórios digitais.

Por fim, importa registar as "barreiras" à partilha de dados. São poucos os investigadores que dispõem dos conhecimentos técnicos indispensáveis à disponibilização dos dados recolhidos de modo a que sejam acedidos por outros utilizadores. No momento do arranque, mesmo aqueles que denotam preocupações com o futuro dos dados que estão prestes a recolher não sabem como actuar e, muitas vezes, a quem se dirigir para obtenção de esclarecimentos. Poucos são os investigadores com a experiência e o domínio de competências técnicas para a organização e o processamento dos dados, embora a vertente mais ignorada seja, ainda, a da criação de metadados capazes de garantir o acesso aos dados durante o seu ciclo de vida, bem como de assegurar a sua preservação a longo termo. Mesmo quando se trata de domínios científicos afins, os modelos, as metodologias, as tecnologias e as ferramentas adoptados pelos investigadores na gestão dos dados não são normalizados, nem obedecem a quaisquer requisitos prévios de normalização.

Para concluir, acrescente-se o receio das despesas resultantes do investimento em gestão de dados, bem como o de não se dispor de tempo suficiente para levar a bom termo a investigação, no caso de se lhe associarem outras tarefas.

3.2 – Infra-estruturas e recursos para a curadoria e partilha de dados

Os repositórios institucionais, quer pelas suas características tecnológicas e organizativas, quer pelo seu número crescente[56], têm sido apontados com frequência como uma das infra-

[56] Registry of Open Access Repositories (ROAR), acessível em: http://roar.eprints.org.

estruturas susceptíveis de dar resposta às necessidades de curadoria de dados resultantes de processos de investigação para as quais não existem outras soluções no imediato. Nesse sentido, a curadoria de dados científicos constitui um repto que pode ser assumido pelos repositórios institucionais e pelos seus administradores. No entanto, subsistem dúvidas que as instituições em moldes individuais possuam o interesse, os recursos e os profissionais com as competências necessárias para lidar com esse desafio[57].

No que concerne à curadoria de dados científicos, o *Research Information Network* (RIN)[58] manifesta também algumas reservas no que concerne a uma solução sustentada apenas em repositórios institucionais. Reportando-se à realidade do Reino Unido, o RIN refere que os *Research Councils UK* (RCUK)[59] apoiam sobretudo centros de dados temáticos e que alguns dos RCUK temem mesmo que o empenhamento institucional esmoreça à medida que terminam avaliações institucionais ou politicas de financiamento em vigor[60]. Outra ressalva secundária manifestada pelos RCUK é a de que um grande número de instituições a desenvolver diferentes sistemas possa traduzir-se em incoerências e em falta de interoperabilidade. No mesmo sentido, Lyon[61] observa que:"*… the growth in institutional repositories as potential recipients of data outputs, means that the landscape is more complex and further thought is required on best practice."*

No que respeita aos serviços de repositórios, num sentido mais lato, deve ser considerada não só a infra-estrutura técnica, como sistemas, normas e protocolos, necessária para garantir a recolha, a preservação e o acesso aos dados, mas também os meios de apoio e suporte do ponto de vista político, legal e ético, relacionados com o acesso e reutilização dos dados científicos para além do contexto inicial em que foram recolhidos.

---

[57] *"No single institution is likely to have the appropriate mix of individuals to maintain and migrate for the future all the data and metadata it has produced in the previous 12 months, let alone over the institution's digital lifetime. It is therefore unlikely that departmental or institutional repositories will be the long term home of academic research data for preservation purposes"*. Heery, R. (2006) Digital Repositories Roadmap: looking forward. Bath: UKOLN. Disponível em: http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/rep-roadmap-v15.pdf [Consultado em 28 de Abril 2010].

[58] O *Research Information Network* (RIN) foi criado em 2005 no Reino Unido por um consórcio que inclui organismos de financiamento do ensino superior, as bibliotecas nacionais e os sete *Research Councils UK* (RCUK). Tem por objectivo apoiar investigadores do Reino Unido, liderando e coordenando iniciativas conducentes à prestação de informação no domínio da investigação científica.

[59] O *Research Councils UK* (RCUK) é uma parceria estratégica entre os sete *Research Councils* do Reino Unido. O RCUK foi criado em 2002 para permitir que os seus membros trabalhassem em conjunto da forma mais eficaz para aumentar o impacto global e a eficácia da sua investigação, formação e inovação, contribuindo dessa forma para a prossecução dos objectivos do Governo do Reino Unido no que concerne à ciência e inovação.

[60] *"…institutional commitment might wane once the Research Assessment Exercise (RAE) is over and pump-priming funding from the Joint Information Systems Committee (JISC) comes to an end."* - RIN (2007) *Research funders' policies for the management of information outputs report*. London: RIN. Disponível em: http://www.rin.ac.uk/system/files/attachments/Research-funders-outputs-report.pdf [Consultado em 2 de Maio 2010].

[61] Lyon, L. (2007) Dealing with data: roles, responsibilities and relationships, Consultancy Report. Bath: UKOLN. Disponível em: http://opus.bath.ac.uk/412/1/dealing_with_data_report-final.pdf [Consultado em 2 de Maio 2010].

De um modo geral, as instituições têm deixado a curadoria de dados científicos a cargo dos designados *data scientists (ou data managers)*, investigadores integrados em grupos de investigação, departamentos ou unidades autónomas, porque consideram que se trata de trabalho especializado. Porém, a curadoria de dados também tem sido conduzida por profissionais na área da informação que laboram em centros de dados ou em bibliotecas, com funções de suporte ou de ligação a departamentos ou disciplinas, daí frequentemente designados por *data librarians*.

O uso de normas adequadas à curadoria dos dados científicos, garantindo a sua preservação e as condições para a sua reutilização é uma tarefa que requer conhecimentos técnicos e disciplinares. Algumas comunidades científicas possuem tradições e práticas no que concerne à curadoria dos dados de maior relevância científica, mas existem outras (pequenos grupos de investigação ou investigadores individuais) onde isso não sucede. Neste sentido, um factor crucial para a implementação de infra-estruturas e recursos para a gestão e partilha de dados passa pela análise e compreensão das metodologias de trabalho dos investigadores e das várias etapas que os dados percorrem em cada área/disciplina.

As universidades parecem também partilhar a preocupação manifestada por Heery[62] relativamente às competências profissionais, porque crêem que são necessários maiores conhecimentos disciplinares para trabalhar com conjuntos de dados científicos do que com publicações científicas. Esta crescente consciencialização da necessidade de competências específicas para a gestão de dados científicos tem conduzido a que algumas instituições encontrem mesmo dificuldades no recrutamento de pessoal qualificado para o efeito.

Sobre este aspecto, Tonge e Morgan[63] acrescentam que é pouco provável que a maioria dos gestores de repositórios institucionais possua o tempo ou as competências profissionais para trabalhar em áreas disciplinares tão específicas e numa única instituição. Por isso, defendem que os progressos neste domínio serão mais facilmente concretizáveis se forem encontradas soluções comuns e partilhadas por múltiplas instituições.

Por outro lado, projectos como o *eBank UK*[64] e *R4L*[65] parecem demonstrar que plataformas de repositórios como o *ePrints*[66] podem ser adaptadas e utilizadas para armazenar dados,

[62] Heery, R. (2006) Digital Repositories Roadmap: looking forward. Bath: UKOLN. Disponível em: http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/rep-roadmap-v15.pdf [Consultado em 28 de Abril 2010].

[63] Tonge, A.; Morgan, P. (2007) Project SPECTRa: JISC Final Report. Cambridge: University of Cambridge Library. Disponível em: http://www.lib.cam.ac.uk/spectra/documents/SPECTRa_Final_Report_v10.doc [Consultado em 2 de Maio 2010].

[64] Sítio do Projecto eBank Uk acessível em: http://www.ukoln.ac.uk/projects/ebank-uk.

[65] Sítio do Projecto R4L acessível em: http://r4l.eprints.org.

eliminando assim a necessidade de cada instituição ter de desenvolver repositórios a partir do zero. Nesse sentido, autores como Coles[67] referem que a questão das competências profissionais poderá não constituir um obstáculo significativo, na medida em que os "tradicionais" gestores de repositórios institucionais podem ser formados e treinados para gerir dados com relativa facilidade desde que os requisitos para a descrição dos mesmos não sejam demasiado exigentes.

No entendimento de Lyon[68], podem ser identificados elementos concorrentes entre os centros de dados e os repositórios institucionais, mas que devem ser postos de lado para se potenciarem relacionamentos mais produtivos e colaborativos. Os centros de dados possuem a experiência e o conhecimento disciplinar essenciais para a gestão de dados, mas, por outro lado, os repositórios institucionais, usualmente implantados e geridos por bibliotecas ou serviços de informação, possuem também uma orientação de "serviço persistente" bastante interiorizada. Os respectivos pontos fortes podem ser utilizados e partilhadas para benefício mútuo.

No desenvolvimento do projecto *Australian Research Repositories Online to the World* (ARROW)[69], por exemplo, a visão original apontava para um único repositório abrangendo um conjunto de diferentes tipos de conteúdos para apoiar as várias esferas de actividade das instituições participantes no projecto: investigação, ensino e administração. Porém, ao colocar-se esta ideia em prática, comprovou-se que se tratava de uma daquelas propostas simples, óbvias mas erradas. Em teoria, a solução de um único repositório seria implementável, mas porventura mais complexa porque a diversidade de objectos passíveis de serem depositados num repositório pode ser muito grande. Acresce que as características de gestão e de acesso a esses objectos são também diferenciadas. Essas diferenças não são facilmente acomodadas na infra-estrutura de um repositório comum e são ainda mais notórias quando se passa de um repositório de publicações para um repositório que também pode conter grandes quantidades e diversidade de dados científicos.

Na sua abordagem aos repositórios de dados, a universidade australiana de Monash[70], para lidar com a questão da preservação de dados científicos e da curadoria de dados em todo o

---

[66] Mais informações sobre a plataforma ePrints disponível em: http://www.eprints.org/software.

[67] Coles, S. (2007). The Repository for the Laboratory (R4L) Project. D-Lib Magazine [online], 13 (3/4). Disponível em: http://www.dlib.org/dlib/march07/03inbrief.html#COLES [Consultado em 4 de Maio de 2010].

[68] Idem.

[69] Sítio do projecto Australian Research Repositories Online to the World (ARROW) acessivel em: http://arrow.edu.au.

[70] Sítio da Universidade de Monash acessível em: http://www.monash.edu.au.

ciclo de investigação, introduziu o conceito de dimensão, de armazenamento de dados e de fronteiras de curadoria para fundar o que designaram por *Data Curation Continua*. Nesta metodologia, as dimensões da gestão da informação foram quantificadas de forma ampla para possuírem valores específicos (com excepção do tempo). Uma análise do espaço de gestão dos dados científicos, baseada em necessidades de utilizadores, revisão da literatura e no trabalho realizado no projecto *Dataset Acquisition, Accessibility and Annotations e-Research Technologies* (DART)[71], mostrou que poderia ser mais difícil identificar valores específicos ao longo de cada dimensão e, em alternativa, a proposta de Monash baseia-se na formação de um *continuum* entre dois pontos. Analisa-se em seguida com maior detalhe alguns desses *continua*.

De acordo com esta metodologia, no que se refere, por exemplo, aos metadados num extremo do *continuum* os dados poderão conter os metadados mínimos imprescindíveis aos criadores e utilizadores: um conjunto de metadados descritivos simples (nome de ficheiro, criador, etc.) e de metadados técnicos específicos da disciplina. No outro extremo do *continuum*, os objectos conterão metadados muito mais ricos, tais como os de proveniência, para efeitos de inteligibilidade, de preservação e para garantir a própria curadoria de dados.

O extremo de um *continuum* pode também descrever repositórios com muitos itens, resultantes de diferentes versões de conjuntos de dados, de experiências fracassadas ou inconclusivas ou de objectos passíveis de eliminação. O ponto oposto do *continuum* serve para descrever repositórios com menos itens. Um número menor de itens resulta, por exemplo, de uma política institucional de gestão de dados ou de os objectos serem referenciados numa dada publicação.

A dimensão dos objectos é outro *continuum* possível e uma questão crítica neste domínio. É normal armazenarem-se nos repositórios objectos com grande diversidade de dimensões. Hoje em dia, grande parte dos projectos de investigação lida de forma regular com objectos de grandes dimensões (ou seja, multi-*gigabytes*) e complexidade. Outros repositórios, por exemplo, repositórios institucionais, muito focados em publicações científicas, são concebidos para operar com objectos de menores dimensões (na ordem dos *megabytes*).

Outro aspecto a considerar é a persistência dos objectos. Os investigadores pretendem com frequência que os dados científicos sejam actualizados à medida que o projecto de investigação progride. Um bom exemplo é um registo de dados climáticos que cresce à

[71] Projecto Dataset Acquisition, Accessibility and Annotations e-Research Technologies (DART) acessível em: http://dart.edu.au.

medida que novos dados são recolhidos. Num objecto de dados, estas mudanças não são adequadas se este estiver interligado a uma publicação como parte de um registo académico permanente. Para este efeito, é preferível ter uma imagem do objecto de dados ou um subconjunto dele extraído.

As próprias práticas de curadoria de dados, ou seja, quem é o responsável pela gestão dos dados (ou que tem controlo sobre eles), são também caracterizáveis por um *continuum*. Os dados científicos disponíveis em espaços colaborativos são muitas vezes geridos pelos próprios investigadores membros da equipa de investigação. Estes gestores podem não possuir as competências ou um mandato para a curadoria de dados a longo prazo. Por esse motivo, no outro extremo deste *continuum* é contemplada a curadoria conduzida por um grupo dedicado dentro da própria instituição.

Um dos processos que pode também ser avaliado no que diz respeito a objectos de dados é o grau de utilização de práticas de preservação. No caso de um grupo de investigação é muito provável que o horizonte de preservação se prolongue para além do término do projecto ou do seu próprio financiamento, mas uma instituição pode assumir horizontes de preservação ainda mais longos, que se ligam por exemplo a razões como a preservação da memória académica e científica ou a obrigações legais assumidas pela instituição.

O trabalho dos investigadores do projecto DART evidenciou igualmente que muitos investigadores são bastante conservadores no que concerne à cedência de dados científicos. Esta preocupação parece estar associada à crescente concorrência pela captação de financiamento para investigação ou na aceitação de artigos científicos. Por esse motivo, muitos investigadores pretendem possuir alguma forma de controlo sobre o acesso aos seus dados antes da publicação. Num repositório de acesso livre, teoricamente, é possível fornecer aos investigadores alguns níveis de controlo em termos do acesso, mas a utilização de repositórios diferentes de acordo com o tipo de acesso (livre ou restrito) aos dados que recolhem é uma alternativa possível.

A franquia do acesso por si só não é razão suficiente para garantir os benefícios do acesso livre. Os objectos depositados nos repositórios também necessitam de visibilidade e de serem pesquisáveis. Num extremo deste *continuum*, há pouca ou nenhuma exposição a *harvesters*[72], o que significa que os objectos de dados com restrições de acesso não são susceptíveis de serem descobertos. No outro extremo deste *continuum*, os conteúdos dos

[72] Um *harvester* é uma aplicação cliente que envia pedidos baseados no protocolo OAI-PMH e geralmente é utilizado por um fornecedor de serviços para recolher de forma automática metadados de repositórios em estádios de produção.

repositórios são expostos a *harvesters* para potenciar a máxima acessibilidade e visibilidade.

Pelo que se observa, a abordagem à constituição de um único repositório sendo inicialmente atractiva, fica desde logo condicionada por vários desafios na fase de implementação. A partilha e curadoria de dados constituem também um repto aos investigadores, às suas instituições, organismos de financiamento, bem como às próprias comunidades científicas no sentido de se alcançarem soluções comuns, fiáveis e reutilizáveis.

3.3 – Enquadramento ético e legal da partilha e curadoria de dados

A disponibilização na web de dados científicos, tendo em vista a sua divulgação para fins de partilha e de reutilização por investigadores, pressupõe que os dados possam ser acedidos pelos interessados. É, porém, do conhecimento geral, que nem todos os dados recolhidos no decurso de processos de investigação são susceptíveis de serem comunicados e acedidos de forma livre e irrestrita. Este aspecto foi, já, abordado de forma sumária na Secção 3.1, pois mais não é do que uma das limitações à disponibilização, em acesso aberto, dos dados recolhidos, produzidos e acumulados pelos investigadores. De facto, o que, por vezes, é interpretado como *resistência* à comunicação e partilha dos resultados de pesquisas pode significar apenas uma preocupação legítima com o sigilo intrínseco a determinados dados, pelo menos durante certos períodos de tempo. Como também já foi referido, acontece com frequência que os investigadores, desconhecendo ou temendo desconhecer os aspectos legais que podem condicionar a partilha e a publicação de dados, optam, por motivos de segurança, por tratar de modo idêntico os dados que são, de facto, confidenciais, e outros que o não são.

Grande parte dos dados recolhidos por investigadores são susceptíveis de ser comunicados, se não na *web*, pelo menos no seio de determinados grupos de interessados, cuja validação, para acesso, seja assegurada pelos gestores de repositórios. Outros dados estão protegidos por lei, outros, ainda, requerem permissões especiais para acesso ou utilização, para além daqueles dados que não podem de forma alguma tornar-se públicos durante determinados períodos de tempo. A confidencialidade, fundamento da segurança da informação, é a única forma de garantir que a informação apenas se encontra disponível para quem estiver autorizado a aceder-lhe[73].

[73] Organização Internacional de Normalização (ISO) – Norma ISSO-17799.

Por todos estes motivos torna-se imprescindível que, aquando do depósito dos dados, os seus criadores documentem o processo de transferência de custódia com toda a informação que regulamenta o acesso e a difusão. As permissões de *copyright*, mas também os direitos de propriedade dos materiais recolhidos, deverão ser definidos previamente ao seu depósito. Este aspecto tem particular relevância quando se trata de dados produzidos a partir de fontes diversificadas e sempre que a investigação tenha sido financiada por mais do que uma organização. Nestes casos, terá que ser assegurada a autorização de todas as instituições intervenientes para depositar e disseminar os dados recolhidos e essas autorizações deverão ser arquivadas em conjunto com os dados.

As questões de *copyright* são mais pertinentes na fase de recolha de dados do que na sua reutilização. Os investigadores das áreas das ciências sociais são particularmente sensíveis a estes aspectos. Sobretudo quando a investigação assenta em entrevistas, questionários e inquéritos, os direitos de *copyright* têm de ser bem esclarecidos. Os entrevistados, por exemplo, detêm a propriedade intelectual dos dados transmitidos por via oral no caso de as transcrições das entrevistas utilizarem uma parte substancial desses dados.

Em projectos que envolvam investigadores ou instituições de vários países, a diversidade das legislações nacionais referentes à propriedade intelectual constitui um aspecto que pode aumentar a complexidade e a dificuldade na definição regras claras no que respeita ao acesso e à utilização dos dados.

Em certas circunstâncias, mesmo os dados confidenciais ou protegidos por outro tipo de reserva podem ser objecto de consulta. Existem dados que podem ser tornados anónimos, embora isso implique a remoção de toda e qualquer informação que permita identificar a quem dizem respeito ou que seja por eles referenciada. Por norma, trata-se de nomes de pessoas e de lugares. O acesso pode, ainda, ser viabilizado mediante a obtenção de uma autorização expressa das entidades a que respeitam. Por fim, e no caso de se tratar de uma operação exequível, poderão encobrir-se informações pessoais relativas aos titulares dos dados, como os seus nomes reais, as moradas, os contactos. Seja como for, pressupõe-se uma utilização cuidada dos dados acedidos em circunstâncias como estas, devidamente acompanhada de uma declaração que a formalize. As boas práticas nesta matéria encaminham para a identificação dos dados que são de acesso público e dos que o não são, numa fase prévia à da sua integração em repositórios digitais.

Como já foi referido atrás, existem diversas iniciativas que têm por objectivo a promoção de soluções jurídicas que facilitem a abertura dos dados científicos. A *Open Data Commons*[74], por exemplo, apresentou uma licença inovadora, a *Public Domain Dedication and License (PDDL)*[75], que possibilita partilhar livremente, modificar e utilizar dados para qualquer finalidade e sem quaisquer restrições. O uso desta licença adequa-se em particular aos bancos de dados ou aos seus conteúdos ("dados"), quer em conjunto quer individualmente.

Na mesma área de actuação o projecto *Science Commons*[76] *(SC)* pretende, entre outros fins, disponibilizar aos investigadores ferramentas que simplifiquem a especificação dos termos sob os quais pretendem partilhar os dados que produzem. Esta iniciativa extrapola a utilização de licenças *CC* no domínio da ciência, encorajando a inovação científica através do acesso de cientistas, universidades e diferentes indústrias aos dados primários e à literatura científica.

As questões jurídicas que se levantam no domínio dos dados científicos não serão, obviamente, resolvidas com a simples adopção de modelos de licenciamento alternativos. Não obstante, a crescente utilização de licenças do tipo *Open Data Commons* ou *Science Commons* poderão contribuir de forma significativa para facilitar a criação, difusão e utilização do conhecimento científico e fomentar novos processos de inovação.

Durante a última década tem-se assistido à multiplicação de iniciativas enquadradas no Movimento de Acesso Livre ao Conhecimento Científico (*Open Access Initiative*). Estas iniciativas têm-se materializado na criação e enriquecimento de repositórios digitais que facultam o acesso aberto à literatura científica produzida por comunidades académicas. Pese embora a disponibilização destas publicações em plataformas *web* significar para os seus utilizadores a possibilidade de aceder à versão integral de trabalhos científicos, bem como a de os descarregar para computadores pessoais, copiar, distribuir e imprimir, os conceitos de *Open Access* e de *full-tex*" não implicam a publicação em acesso aberto dos dados brutos, subjacentes aos processos de investigação. Registe-se, porém, que o panorama internacional patenteia o aumento da frequência de publicação dos resultados de investigação simultaneamente com a dos dados primários que lhe são subjacentes. Existem instituições que, a nível internacional, financiam projectos de investigação e são ao mesmo tempo autoras de mandatos de *Open Access* e de *Open Data*.

---

[74] Mais informações sobre a *Open Data Commons* acessíveis em: http://www.opendatacommons.org/.

[75] Mais informações em sobre a licença *Public Domain Dedication and License (PDDL)* acessíveis em: http://www.opendatacommons.org/licenses/pddl/1-0/.

[76] Mais informações sobre a *Science Commons* acessíveis em: http://sciencecommons.org.

4 – CONCLUSÕES

O presente relatório revela dois aspectos essenciais quanto ao tema da curadoria e da partilha dos dados científicos, nomeadamente através de repositórios.

Em primeiro lugar, constatou-se o crescente interesse que esta área tem despertado na comunidade científica e nas instituições onde se realiza investigação. De facto, em especial nos últimos cinco anos, têm-se multiplicado as actividades, iniciativas, projectos e serviços relacionados com a curadoria e partilha de dados científicos. Essas iniciativas têm tido origens e âmbitos muito diversos, desde comunidades informais de investigadores até grandes projectos internacionais, passando por inúmeras iniciativas e serviços institucionais. Apesar de muitos investigadores ainda estarem pouco conscientes e sensibilizados para as questões relacionadas com os dados científicos e de em muitas instituições ou comunidades disciplinares as actividades neste domínio ainda serem quase inexistentes ou muito incipientes, a evolução registada nos últimos anos tem sido muito significativa e prenuncia uma atenção mais generalizada e interiorizada entre os vários actores da investigação científica.

Em segundo lugar, constatou-se também que, apesar de algumas excepções em áreas científicas específicas onde já existem actividades neste domínio há várias décadas, a curadoria e partilha de dados científicos é uma área "jovem", ainda em formação e consolidação. Como a maioria dos serviços e projectos é recente, a maturidade, escalabilidade e sustentabilidade de diversas soluções tecnológicas e organizativas está ainda por demonstrar, pelo menos de forma integral. A "juventude" da curadoria e partilha dos dados científicos é uma oportunidade para a investigação, desenvolvimento e teste de novos serviços e tecnologias, mas constitui simultaneamente um risco para o investimento na criação e oferta de serviços de qualidade profissional.

Para além destes dois aspectos, o estudo realizado permite ainda concluir que a curadoria dos dados científicas, para ser verdadeiramente efectiva e sustentável, exige a participação de todas as partes (os investigadores, as instituições onde trabalham e os organismos de financiamento) envolvidas na produção dos dados e no processo de investigação.

Por isso mesmo, sugerem-se, de seguida, algumas acções, atitudes e orientações que poderão ser desenvolvidas, pelos diferentes actores neste processo, no sentido de promover e facilitar o processo de curadoria e partilha de dados em Portugal.

4.1 – Investigadores

Os investigadores são os protagonistas centrais de qualquer estratégia de curadoria e partilha dos dados científicos, uma vez que o destino dos dados é em grande medida decidido no momento da sua criação e recolha, ou mesmo antes de estas se iniciarem, durante o planeamento da investigação.

Nesse sentido, será importante que os investigadores:

1. Incluam a curadoria, e eventual partilha, dos dados científicos, no processo de planeamento da investigação, analisando e identificando as soluções e as práticas adequadas, desejavelmente produzindo um plano de curadoria que abranja, entre outros, os seguintes aspectos:
    a) Descrição e identificação dos conjuntos de dados a produzir na investigação, através de metadados apropriados e, quando possível, normalizados;
    b) Definição do ciclo de vida dos dados recolhidos na investigação, identificando os prazos associados ao depósito em infra-estruturas "externas" (repositórios), à preservação (temporária ou permanente) e, se aplicável, ao acesso aberto e partilha dos dados;
    c) Identificação dos repositórios ou outras infra-estruturas onde os dados deverão ser depositados, a fim de poderem ser preservados e, se aplicável, disponibilizados e partilhados;
    d) Identificação dos condicionalismos éticos e legais, nomeadamente quanto à confidencialidade, protecção de dados pessoais, direitos de autor e propriedade intelectual, que poderão estar associados aos dados que vão ser produzidos, recolhidos ou processados, e que poderão exigir medidas e soluções especiais;
    e) Análise e identificação dos custos, incluindo o tempo de trabalho, associados ao processo de curadoria dos dados.
2. Procurem colaborar e cooperar com os serviços e projectos de curadoria de dados relevantes para o seu contexto, disciplinar e institucional, a fim de conhecer, utilizar e promover as boas práticas neste domínio.
3. Divulguem e partilhem os dados científicos que produzam, tão cedo e tão amplamente quanto possível em cada caso, sem prejuízo dos seus próprios interesses ou dos constrangimentos legais e éticos que possam existir.

4.2 – Instituições de investigação

A generalidade dos dados científicos e dos processos de investigação que lhes dão origem realiza-se num contexto institucional. As instituições que realizam investigação, sejam universidades, centros de investigação e laboratórios públicos, ou organizações privadas, podem desempenhar um papel crucial na curadoria dos dados científicos, quer disponibilizando e mantendo infra-estruturas, quer definindo políticas e procedimentos.

Nesse sentido, constituirão elementos centrais de sucesso para a curadoria dos dados que as instituições de investigação:

1. Realizem um recenseamento e diagnóstico da situação existente no que diz respeito aos dados científicos que produzem. Para além da identificação dos dados existentes, deverá ser avaliado o estado actual de curadoria e as práticas de recolha, processamento, preservação e acesso a esses dados.
2. Disponibilizem, de forma autónoma ou contribuindo para a sua criação de forma colaborativa e partilhada com outras instituições, infra-estruturas e serviços para a curadoria dos dados científicos, seja através dos repositórios institucionais já existentes, seja através de repositórios de dados concebidos para esse efeito.
3. Atribuam a competência pela área da curadoria de dados científicos a uma unidade organizacional da instituição, devidamente apetrechada com os recursos financeiros e humanos necessários ao desempenho dessa função.
4. Incentivem os investigadores a preocupar-se com a curadoria dos dados que produzem, a depositá-los no(s) repositório(s) institucionalmente adequados e a partilhá-los sempre e logo que possível. Neste sentido, poderá ser útil vir a considerar:
    a) Acções de sensibilização e divulgação;
    b) Produção de documentação e divulgação de casos de estudo, boas práticas e exemplos de sucesso na curadoria e partilha de dados científicos;
    c) Oferta de serviços de aconselhamento, consultoria e formação sobre os aspectos técnicos e legais;
    d) Produção e divulgação de políticas, orientações, directrizes, normas e procedimentos relativos à curadoria de dados que os investigadores devam utilizar durante o ciclo de vida da investigação.
5. Definam políticas institucionais que induzam o depósito dos dados científicos em repositórios institucionalmente adequados, estimulem a partilha dos dados depositados tão cedo e de forma tão ampla quanto seja possível, e reconheçam e

premeiem os investigadores que cumpram os requisitos de curadoria dos dados definidos pelas instituições.

6. Avaliem e identifiquem as necessidades de formação de técnicos de curadoria de dados para as diferentes áreas científicas e disciplinares e, no caso das instituições de ensino superior, considerem a oferta de formação adequada para responder a essas necessidades.

4.3 – Organismos financiadores de investigação

Os organismos, públicos e privados, que financiam a investigação científica em Portugal podem ter um papel activo na promoção de boas práticas de curadoria e acesso aos dados científicos já que, como financiadores, detêm um poder efectivo quer para a definição e implementação das políticas necessárias, quer para o apoio à investigação e ao funcionamento de infra-estruturas no domínio da curadoria de dados.

Nesse sentido, a acção dos organismos financiadores pode tornar-se um importante contributo para o incremento da curadoria e partilha de dados científicos se:

1. Definirem políticas que exijam, ou pelo menos valorizem de forma significativa na avaliação para financiamento, a existência de um plano de curadoria de dados nos projectos de investigação que financiem.
2. Definirem políticas e procedimentos que exijam, com as excepções necessárias e as limitações adequadas, o depósito dos dados científicos em repositórios sustentáveis e o acesso aberto a esses dados sempre e logo que possível.
3. Considerarem como elegíveis para financiamento, nos projectos que financiam, as eventuais despesas dos investigadores com actividades de curadoria e partilha de dados.
4. Disponibilizarem financiamento específico para a realização de projectos de investigação e desenvolvimento no domínio da curadoria de dados científicos, independentemente das áreas científicas de origem dos dados.

4.4 – Responsáveis por repositórios

Os responsáveis pelos repositórios que albergam, ou poderão albergar, dados científicos desempenham um papel crítico no processo de curadoria dos dados. Apesar da juventude e "imaturidade" das infra-estruturas e tecnologias que se têm desenvolvido nos últimos anos, é necessário assegurar serviços de qualidade e, em especial, a preservação dos dados que, no presente momento, estão já confiados aos repositórios.

Assim, é muito importante que os responsáveis dos repositórios desenvolvam acções no sentido de:

1. Assegurar que os repositórios colocados à disposição dos investigadores sejam infra-estruturas robustas e fiáveis do ponto de vista tecnológico e da segurança da informação;
2. Disponibilizar aos investigadores serviços e ferramentas de apoio à curadoria e partilha de dados, tais como:
    a) Modelos de protocolos e licenças que os investigadores e produtores dos dados possam usar para regular o arquivo dos dados nos repositórios (definindo as condições de depósito, preservação e acesso, incluindo eventuais condições de reutilização ou existência de restrições por razões de confidencialidade);
    b) Acções e materiais de sensibilização e formação;
    c) Serviços de aconselhamento jurídico no que respeita aos dados científicos, nomeadamente quanto à propriedade intelectual e à salvaguarda da confidencialidade e privacidade;
3. Recolher e disponibilizar informação sobre o acesso e a utilização dos conjuntos de dados que gerem, comprometendo-se a fornecer informação estatística aos investigadores que os recolheram e processaram.
4. Acompanhar as iniciativas internacionais relacionadas com a criação, gestão e manutenção de repositórios de dados, recolhendo a experiência de projectos pioneiros e aprendendo com os seus resultados.

♦

A materialização das linhas de acção acima identificadas, para os diferentes participantes dos processos de investigação, produção de dados científicos e curadoria de dados, pode ser apoiada e incentivada através de actividades e iniciativas do projecto Repositório Científico de Acesso Aberto de Portugal.

Os dados científicos são, a par da literatura científica, um dos resultados da investigação. Apesar de grandes diferenças nos requisitos, problemas e soluções entre o arquivo e a disponibilização de publicações e a curadoria e partilha de dados, existem obviamente inúmeros pontos comuns.

Assim, a experiência acumulada pelo projecto RCAAP no domínio dos repositórios institucionais pode constituir uma base para o lançamento de iniciativas na área dos repositórios de dados científicos.

GLOSSÁRIO

Termo	Descrição
Acesso Aberto (*Open Access*)	Também designado *Acesso Livre* refere-se ao acesso à literatura científica em formato electrónico, em particular aos artigos de revistas com revisão por pares, sem barreiras de preço e de permissões. Os três principais documentos definidores do Acesso Aberto são as Declarações de *Budapeste*, *Bethesda* e *Berlim*.

Termo	Descrição
Base de Dados *(Database)*	Uma base de dados é uma estrutura desenhada para o armazenamento e a interrogação de grandes volumes de dados. A base de dados é suportada por um modelo de dados que descreve as entidades representadas, os respectivos atributos e as relações entre elas. Tipicamente um modelo de base de dados é desenhado para um domínio e tendo em vista um conjunto de aplicações, e capta as entidades relevantes para estas no domínio.

Termo	Descrição
Conjunto de Dados *(Dataset)*	Um conjunto de dados, ou conjunto de dados científicos, reúne itens de natureza semelhante que foram coleccionados para servirem de base a investigação. Nos dados científicos há tipicamente uma granularidade bem estabelecida. Se os dados são registos de entrevistas, cada entrevista é um item, com a sua estrutura de perguntas, e o conjunto de dados é o conjunto das entrevistas (e não o conjunto das respostas a perguntas individuais). A granularidade pode ser óbvia, como no caso de um conjunto de fotografias, ou dependente da visão dos investigadores sobre a realidade, como acontece num conjunto de dados botânicos, em que a unidade pode ser a espécie ou o exemplar dessa espécie que vai ser estudado.

Termo	Descrição
Curadoria de Dados *(Data Curation)*	A curadoria de dados designa o conjunto de acções que garantem que um conjunto de dados é genuíno, permitindo o seu uso por outros que não os seus produtores. A curadoria pode envolver acções de descrição dos dados, de ligação destes a outros que os tornem inteligíveis, de registo dos usos que tenham e dos resultados a que tenham dado origem. A curadoria envolve também acções de preservação, em que a representação dos dados e os seus metadados tenham de ser modificados. As acções de curadoria e de gestão de dados têm alguma intersecção, sendo as de gestão mais independentes do conteúdo e do uso.

Termo	Descrição
Dados Científicos *(Research Data)*	Dados que são produzidos no contexto de investigação científica ou que de alguma forma são usados em investigação. Estes dados podem ser criados para efeito de processamento científico, como nos dados atmosféricos usados para previsão meteorológica, ou os dados recolhidos de sensores para monitorar o estado de um edifício. Há dados que são obtidos como resultados do processamento automático de objectos (eles próprios representados como dados), como, por exemplo, os histogramas de cor obtidos através do processamento de uma colecção de imagens. Há ainda dados que não sendo produzidos para investigações acabam por ser objecto dela, como as contribuições que os utilizadores de uma rede social fazem, na forma de textos ou outros conteúdos, e que acabam por ser utilizados para estudos sociológicos.

Termo	Descrição
Gestão de dados *(Data Management)*	Na gestão de dados científicos incluem-se as acções de representação dos dados e seu armazenamento, a associação de metadados que os descrevem, que ajudam a interpretá-los e documentam o seu uso, a organização dos dados em colecções, a indexação dos dados para pesquisa e todas as formas de apresentação dos dados. A gestão de dados pode ser realizada por pessoas que não têm conhecimento directo do seu significado, desde que os dados tenham sido descritos de forma completa pelos seus produtores.

Termo	Descrição
Metadados *(Metadata)*	Os metadados são dados cujo propósito é garantir a autenticidade, descrever, tornar acessíveis, ou de alguma forma qualificar e aumentar a inteligibilidade dos dados de base. Quando o significado dos dados está acessível ao utilizador menos especializado, como é o caso dos registos de publicações, existem normas bem estabelecidas para os metadados, neste exemplo as diversas normas bibliográficas. Na maioria das áreas que produzem dados científicos, não estão fixadas normas para metadados, o que torna a tarefa de descrição um desafio substancial.

Termo	Descrição
Plataforma digital de investigação *(e-research)*	Uma plataforma digital de suporte à investigação fornece tecnologias para suportar o processo de investigação, incluindo a colaboração entre grupos, a recolha de dados, a sua análise, a publicação de resultados, o armazenamento e a partilha dos dados. As tecnologias disponíveis neste domínio incluem os ambientes de investigação virtuais, a computação *grid*, os serviços de visualização, a mineração de dados.

Termo	Descrição
Repositórios institucionais *(Institutional Repositoires)*	Um repositório institucional é uma infra-estrutura mantida por uma organização, tal como uma universidade ou um centro de investigação, com o propósito de coleccionar e preservar a sua produção científica, técnica ou administrativa, e de lhe dar visibilidade. Os repositórios institucionais podem contribuir para iniciativas mais alargadas, como a agregação da publicação científica a nível nacional. Os repositórios têm também uma função importante na auditoria das organizações.

Termo	Descrição
Serviços de Repositório *(Repository Services)*	Serviços oferecidos por uma infra-estrutura técnica que permite o armazenamento, o acesso, a descrição, a disseminação e a preservação de objectos digitais. Nesta infra-estrutura também estão presentes os serviços de apoio aos investigadores em aspectos técnicos, legais e das políticas de criação, do depósito e partilha de resultados de investigação.

BIBLIOGRAFIA

BURTON, A., & TRELOAR, A. (2009). Publish My Data: A composition of services from ANDS and ARCS. eScience, p. 164-170. Fifth IEEE International Conference on eScience. Disponível em: http://doi.ieeecomputersociety.org/10.1109/e-Science.2009.31 [Consultado em 8 de Maio 2010].

BURTON, A., & TRELOAR, A. (2010). Publish My Data: the design and implementation of a loosely-coupled data 'publishing' service. Proceedings of VALA 2010. Melbourne, Australia. Disponível em: http://www.vala.org.au/vala2010/papers2010/VALA2010_123_Burton_Final.pdf

COLES, S. (2007). The Repository for the Laboratory (R4L) Project. D-Lib Magazine [online], 13 (3/4). Disponível em: http://www.dlib.org/dlib/march07/03inbrief.html#COLES [Consultado em 4 de Maio de 2010].

FRY, J., LOCKYER, S., OPPENHEIM, C., HOUGHTON, J., & RASMUSSEN, B. (2008). Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes". Loughborough University, Centre for Strategic Economic Studies. Disponível em: http://hdl.handle.net/2134/4600

GOLD, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. D-Lib Magazine [online], 13 (9/10). Disponível em: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html [Consultado em 4 de Maio 2010].

GOLD, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 2. Libraries and the Data Challenge: Roles and Actions for Libraries. D-Lib Magazine [online], 13 (9/10). Disponível em: http://www.dlib.org/dlib/september07/gold/09gold-pt2.html [Consultado em 4 de Maio 2010].

GREEN, A., & GUTMAN, M. P. (2007). Building partnerships among Social Science researchers, institution-based repositories, and domain specific data archives. Dublin, Ohio: OCLC Systems and Services: International Digital Library Perspectives, 23 (1). Preprint disponível em: http://hdl.handle.net/2027.42/41214

GREEN, A., MACDONALD, S., & RICE, R. (2009). Policy-making for Research Data in Repositories: A Guide. London: JISC funded DISC-UK Share Project. Disponível em: http://www.disc-uk.org/docs/guide.pdf

Heery, R. (2006). Digital Repositories Roadmap: looking forward. Bath: UKOLN. Disponível em: http://www.ukoln.ac.uk/repositories/publications/roadmap-200604/rep-roadmap-v15.pdf [Consultado em 28 de Abril 2010].

HEY, Tony; TRANSLEY, Stewart; TOLLE, Kristin, eds. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, Washington, Microsoft Research.

KEY PERSPECTIVES. (2010), *"Data dimensions: disciplinary differences in research data sharing, reuse and long term viability: A comparative review based on sixteen case studies"*. DCC SCARP Synthesis Report comissioned by the Digital Curation Centre. Disponível em: http://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf [consultado em 8 de Maio de 2010].

LYON, L. (2007) Dealing with data: roles, responsibilities and relationships, Consultancy Report. Bath: UKOLN. Disponível em: http://opus.bath.ac.uk/412/1/dealing_with_data_report-final.pdf [Consultado em 4 de Maio 2010].

LYON, L. (2009). Open Science at Web-Scale: Optimising Participation and Predictive Potential Consultative Report. DCC Report comissioned by.JISC. Disponível em: http://www.jisc.ac.uk/media/documents/publications/research/2009/open-science-report-6nov09-final-sentojisc.pdf

MACDONALD, S. (2009). Edinburgh DataShare - A DSpace Data Repository: Achievements and Aspirations. Apresentação efectuada em *Fedora UK&I&EU Meeting - Fedora-based e-Research environments.* University of Oxford, 8 December 2009.. Disponível em: http://hdl.handle.net/1842/3201

MARTINEZ-URIBE, L. (2007). Digital Repository Services for Managing Research Data: What Do Oxford Researchers Need? IASSIST Quarterly, 31 (3/4), Fall & Winter 2007.

OECD (2004). OECD: Declaration on Access to Research Data From Public Funding. Paris: OECD. Disponível em: http://www.oecd.org/document/15/0,3343,en_2649_34487_25998799_1_1_1_1,00.html [Consultado em 26 de Abril 2010]

OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. Paris: OECD. Disponível em: http://www.oecd.org/dataoecd/9/61/38500813.pdf [consultado em: 26 de Abril de 2010]

PIWOWAR, H. A., & CHAPMAN, W. W. (2010). Public sharing of research datasets: A pilot study of associations. Journal of Informetrics, Vol. 4, No. 2. (April 2010), pp. 148-156.

RICE, R. (2009). *DISC-UK DataShare Project: Final Report.* Edinburgh: University of Edinburgh, April 2009. Disponível em: http://ie-repository.jisc.ac.uk/336/1/DataSharefinalreport.pdf

RICE, R. (2009b). Open Data and Institutional Repositories [Presentation]. 4th Conference on Open Access. University of Minho, Braga, Portugal: 26-27 November, 2009.

RIN (2007) Research funders' policies for the management of information outputs report. London: Research Information Network. Disponível em: http://www.rin.ac.uk/system/files/attachments/Research-funders-outputs-report.pdf [Consultado em 2 de Maio 2010].

RIN (2007b) Stewardship of digital research data: a framework of principles and guidelines (Draft for consultation). London: RIN. Disponível em: http://www.rin.ac.uk/system/files/attachments/Stewardship-data-guidelines.pdf [Consultado em 2 de Maio 2010].

RUUSALEPP, R. (2008). Comparative Study of International Approaches to Enabling the Sharing of Research Data. DCC Report comissioned by.JISC. Disponível em: http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/Data-Sharing-Report.pdf

SESINK, L., VAN HORIK, R., & HARMSEN, H. (2008). Data Seal of Approval: Quality Guidelines for Digital Research Data in the Netherland. The Hague: Data Archiving and Networked Services. The Hague, Data Archiving and Networked Services - 2nd ed. DANS, 2010. ISBN 978 9490 531 027.

SWAN, A., & BROWN, S. (2008). To Share or not to Share: Publication and Quality Assurance of Research Data Outputs. London: Research Information Network.

SWAN, A. & BROWN, S. (2008) The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Technical Report, School of Electronics & Computer Science, University of Southampton. Disponível em: http://eprints.ecs.soton.ac.uk/16675 [Consultado em 11 de Maio 2010].

TRELOAR, A., GROENEWEGEN, D., & HARBOE-REE, C. (2007). The Data Curation Continuum: Managing Data Objects in Institutional Repositories. D-Lib Magazine [online], 13 (9/10). Disponível em: http://www.dlib.org/dlib/september07/treloar/09treloar.html [Consultado em 20 de Abril 2010].

TRELOAR, A., & WIKINSON, R. (2008). Rethinking Metadata Creation and Management in a Data-Driven Research World. eScience, p.782-789. Fourth IEEE International Conference on eScience. Disponível em: http://doi.ieeecomputersociety.org/10.1109/eScience.2008.41 [Consultado em 8 de Maio 2010].

TONGE, A.; MORGAN, P. (2007). Project SPECTRa: JISC Final Report. Cambridge: University of Cambridge Library. Disponível em: http://www.lib.cam.ac.uk/spectra/documents/SPECTRa_Final_Report_v10.doc [Consultado em 4 de Maio 2010].

WILLIAMS, R., PRYOR, G., BRUCE, A., MARSDEN, W., & MACDONALD, S. (2009). Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences. London: Research Information Network.

PROJECTOS

ANDS - Australian National Data Service: http://ands.org.au/

ARROW - Australian Research Repositories Online to the World: http://arrow.edu.au/

CAVA - Human Communication: an Audio-Visual Archive: http://www.ucl.ac.uk/ls/cava/

CESSDA - Council of European Social Science Data Archives: http://www.cessda.org/

CLARIN - Common Language Resources and Technology Infrastructure: http://www.clarin.eu/external/

DAF - Data Audit Framework: http://www.data-audit.eu/

DANS – Data Archiving and Networked Services: http://www.dans.knaw.nl/en

DARIAH - Digital Research Infrastructure for the Arts and Humanities: http://www.dariah.eu/

DART - Dataset Acquisition, Accessibility and Annotations e-Research Technologies: http://dart.edu.au/

eBank Uk: http://www.ukoln.ac.uk/projects/ebank-uk

Google public data explorer: http://www.google.com/publicdata/directory

IVOA - International Virtual Observatory Alliance: http://www.ivoa.net/

Lifespan Initiative for the Research and Data Archive Repository: http://www.lifespancollection.org.uk/

MDC - Materials Data Centre: http://www.materialsdatacentre.com/

METAFOR - Common Metadata for Climate Modelling Digital Repositories: http://metaforclimate.eu/

OAD - Data Repositories: http://oad.simmons.edu/oadwiki/Data_repositories

ODC - Open Data Commons: http://www.opendatacommons.org/

R4L – Repository for the laboratory: http://r4l.eprints.org/

UCSC Genome Browser: http://genome.ucsc.edu/cgi-bin/hgTracks?org=human

UKDA - UK Data Archive: http://www.data-archive.ac.uk